5-2013

Nonparametric Discriminant Analysis in Forensic Ancestry Estimation: An Assessment of Utilitzed and Alternative Statistical Methods

David Anthony Mercer
dmercer2@utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by David Anthony Mercer entitled "Nonparametric Discriminant Analysis in Forensic Ancestry Estimation: An Assessment of Utilitzed and Alternative Statistical Methods." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Arts, with a major in Anthropology.

Amy Z. Mundorff, Major Professor

We have read this thesis and recommend its acceptance:

William Seaver, Richard Jantz

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Nonparametric Discriminant Analysis in Forensic Ancestry Estimation: An Assessment of Utilized and Alternative Statistical Methods

A Thesis Presented for the

Master of Arts Degree

The University of Tennessee, Knoxville

David Anthony Mercer

May 2013

ACKNOWLEDGEMENTS

I would like to thank my committee members – Dr. Amy Mundorff, Dr. William Seaver, and Dr. Richard Jantz – for their guidance and patience throughout the process of this research as well as their interest in its successful completion. Dr. Mundorff was especially patient and helpful during my search for a research topic. She always provided honest feedback and recommendations with an interest in my success. I would also like to thank The University of Tennessee Forensic Anthropology Center for their provision of the Forensic Data Bank's craniometric. I am very appreciative for the discussions and recommendations provided by Dr. Bridget Algee-Hewitt and Dr. Benjamin Auerbach. While they had no responsibilities to me, they willingly listened to my questions and contributed constructive feedback. Finally, I would like to thank my family, especially my parents, for their support and confidence in me throughout my education.

ABSTRACT

This thesis evaluates and compares the performances of four discriminant analysis techniques in forensic ancestry estimation using craniometric variables. Giles and Elliot (1962) were the first anthropologists to use discriminant analysis for ancestry estimation. They used Linear Discriminant Analysis (LDA) in an attempt to predict American White, American Black, and American Indian ancestry from craniometric variables. LDA has since been the dominant discriminant technique used for this purpose. It is the method that is exclusively used in FORDISC (Ousley and Jantz, 2005) and, until recently, was the only method applied to forensic craniometric ancestry estimation.

LDA, however, assumes the data for each group in the analysis are multivariate normally distributed and the group covariance matrices are equal. These assumptions are not usually addressed in research; they are often assumed as satisfied (Feldesman, 2002). In fact FORDISC includes a test for equal covariances, but not multivariate normality. It assumes the latter condition is met (Ousley and Jantz, 2012). Furthermore, it does not provide an alternative option when LDA's assumptions are violated.

This thesis evaluates and compares the assumptions and performances of LDA and three other discriminant techniques (i.e., quadratic discriminant analysis, k-nearest neighbor analysis, and classification trees) in craniometric ancestry estimation. Each method has unique assumptions about the data, so each may be appropriate for different situations. It is important to apply methods with satisfied assumptions because the results may not be interpretable or gerneralizable otherwise.

The results show that a few outliers are often the cause of violations of multivariate normality. However, covariance equality is difficult to achieve and was not present for any evaluation. LDA had the best overall classification performance. However, its assumptions are often violated. Classification trees are the recommended alternative when LDA's assumptions are not met. Though its performance is likely lower than that of LDA, it offers many advantages that make it a useful method, such as its lack of data assumptions.

Table of Contents

INTRODUCTION **1**

CHAPTER I LITERATURE REVIEW **4**

Early Statistical Explorations of Human Variation **4**

Software for Craniometric Assessment of Ancestry **6**

A Controversy in the Field: Anthropology and Forensic Race Estimation **7**

The Biological Race Concept: Contemporary Views in Anthropology **10**

FORDISC Methodology and Potential Limitations **13**

The Need for Accurate and Reliable Statistical Figures in Forensics **14**

CHAPTER II MATERIALS AND METHODS **16**

Materials **16**

Missing Value Analysis and Imputation **18**

Case Deletion 20

Which Variables and Observations to Retain? 21

Regression Imputation 26

Estimations for this Data 28

Another Considered Imputation Option 29

Methods **30**

Linear Discriminant Analysis 31

Quadratic Discriminant Analysis 32

K-Nearest Neighbor Analysis (KNN) 33

Classification Trees 34

Variable Selection 38

Evaluation and Comparison 40

CHAPTER III RESULTS AND DISCUSSION 43

Results 43

Linear Discriminant Analysis 43

Quadratic Discriminant Analysis 46

K-Nearest Neighbor Analysis 48

Classification Trees 51

Discussion 54

CONCLUSION AND RECOMMENDATIONS 65

LIST OF REFERENCES 71

APPENDIX 76

VITA 128

LIST OF TABLES

Table 1. Sex, ID Status, and Race samples and missing values. 16

Table 2. Group samples. 16

Table 3. ID Statuses by group. 17

Table 4. Group samples using only positive IDs. 17

Table 5. Missing values by variable 19

Table 6. Missingness for the key subset. 23

Table 7 - McNemar results. 25

Table 8. Final Sample. 26

Table 9. Training sample size 42

Table 10. Holdout sample sizes 42

Table 11. LDA classification rates in % 45

Table 12. QDA forward selection results 47

Table 13. QDA classification rates in % 48

Table 14. KNN (*k*=3) forward selection results 49

Table 15. KNN (*k*=3) classification rates in % 50

Table 16. KNN (*k*=1) results in % 51

Table 17. Classification tree results 53

Table 18. Variables selected by the discriminant methods 54

Table 19. Overall classification rates in % 56

Table 20. Group holdout classification rates in % 57

Table 21. Group re-substitution classification rates in % 57

Table 22. 13 variable LDA ancestry aggregated cross-validation classification rate ... 104

Table 23. 13 variable LDA ancestry aggregated holdout classification rate 104
Table 24. 6 variable LDA ancestry aggregated cross-validation classification rate 104
Table 25. 6 variable LDA ancestry aggregated holdout classification rate 105
Table 26. 3 variable QDA ancestry aggregated cross-validation classification rate 105
Table 27. 3 variable QDA ancestry aggregated holdout classification rate 105
Table 28. 6 variable QDA ancestry aggregated cross-validation classification rate 106
Table 29. 6 variable QDA ancestry aggregated holdout classification rate 106
Table 30. 5 variable KNN ancestry aggregated cross-validation classification rate 106
Table 31. 5 variable KNN ancestry aggregated holdout classification rate..... 107
Table 32. 8 variable KNN ancestry aggregated cross-validation classification rate 107
Table 33. 8 variable KNN ancestry aggregated holdout classification rate..... 107
Table 34. 13 variable LDA sex aggregated cross-validation classification rate 108
Table 35. 13 variable LDA sex aggregated holdout classification rate 108
Table 36. 6 variable LDA sex aggregated cross-validation classification rate 108
Table 37. 6 variable LDA sex aggregated holdout classification rate 109
Table 38. 3 variable QDA sex aggregated cross-validation classification rate 109
Table 39. 3 variable QDA sex aggregated holdout classification rate 109
Table 40. 6 variable QDA sex aggregated cross-validation classification rate 110
Table 41. 6 variable QDA sex aggregated holdout classification rate 110
Table 42. 5 variable KNN sex aggregated cross-validation classification rate 110
Table 43. 5 variable KNN sex aggregated holdout classification rate 111
Table 44. 8 variable KNN sex aggregated cross-validation classification rate 111
Table 45. 8 variable KNN sex aggregated holdout classification rate 111

LIST OF FIGURES

Figure 1. Percent of observations by percent missing values 20

Figure 2. Percent of observations by percent of missing values: 12-variable subset. .. 24

Figure 3. SAS stepwise selection results 44

Figure 4. All variables White male c-q plot 78

Figure 5. All variables White female c-q plot 79

Figure 6. All variables Black male c-q plot..... 80

Figure 7. All variables Black female c-q plot..... 81

Figure 8. All variables Hispanic male c-q plot..... 82

Figure 9. 13-variable LDA White male c-q plot..... 83

Figure 10. 13-variable LDA White female c-q plot..... 84

Figure 11. 13-variable LDA Black male c-q plot 85

Figure 12. 13-variable LDA Black female c-q plot 86

Figure 13. 13-variable LDA Hispanic male c-q plot 87

Figure 14. 6-variable LDA White male c-q plot..... 88

Figure 15. 6-variable LDA White female c-q plot..... 89

Figure 16. 6-variable LDA Black male c-q plot 90

Figure 17. 6-variable LDA Black female c-q plot 91

Figure 18. 6-variable LDA Hispanic male c-q plot 92

Figure 19. 3-variable QDA White male c-q plot..... 93

Figure 20. 3-variable QDA White female c-q plot..... 94

Figure 21. 3-variable QDA Black male c-q plot 95

Figure 22. 3-variable QDA Black female c-q plot 96
Figure 23. 3-variable QDA Hispanic male c-q plot 97
Figure 24. 8-variable KNN White male c-q plot 98
Figure 25. 8-variable KNN White female c-q plot 99
Figure 26. 8-variable KNN White female c-q plot 100
Figure 27. 8-variable KNN Black male c-q plot..... 101
Figure 28. 8-variable KNN Black female c-q plot..... 102
Figure 29. 8-variable KNN Hispanic male c-q plot..... 103
Figure 30. 13-variable LDA cross-validation posterior probabilities histogram 112
Figure 31. Descriptive statistics - LDA cross-validation misclassification posterior probabilities 112
Figure 32. 13-variable LDA holdout posterior probabilities histogram 113
Figure 33. Descriptive statistics- LDA holdout misclassification posterior probabilities 113
Figure 34. 3-variable QDA cross-validation posterior probabilities histogram 114
Figure 35. Descriptive statistics - QDA cross-validation misclassification posterior probabilites..... 114
Figure 36. 3-variable QDA holdout posterior probabilities histogram 115
Figure 37. Descriptive statistics - QDA holdout misclassification posterior probabilities 115
Figure 38. Descriptive statistics - LDA re-substitution misses 116
Figure 39. LDA re-substitution misses quartile statistics 116
Figure 40. LDA re-substitution misses histogram..... 117
Figure 41. Descriptive statistics - LDA cross-validation misses..... 118

Figure 42. LDA cross-validation misses quartiles 118
Figure 43. LDA cross-validation misses histogram 119
Figure 44. Descriptive statistics - LDA holdout misses 120
Figure 45. LDA holdout misses quartiles 120
Figure 46. LDA holdout misses histogram 121
Figure 47. Descriptive statistics - QDA resubstitution misses 122
Figure 48. QDA resubstitution misses quartiles 122
Figure 49. QDA resubstitution misses histogram 123
Figure 50. Descriptive statistics - QDA cross-validation misses 124
Figure 51. QDA cross-validation quartiles 124
Figure 52. QDA cross-validation histogram 125
Figure 53. Descriptive statistics - QDA holdout misses 126
Figure 54. QDA holdout misses quartiles 126
Figure 55. QDA holdout misses histogram 127

INTRODUCTION

Metric analysis of population affinity has been a large part of biometric studies throughout the 20th century. Many early statisticians – such as Pearson, Fisher, and Mahalanobis – conducted statistical analyses of population and taxonomic affinity. In fact, Pearson (1926) developed the Coefficient of Racial Likeness specifically for craniologists and physical anthropologists to use in estimating population affinity from craniometric variables. Though Mahalanobis later revealed problems with this statistic (other than the name itself), it played a role in the eventual developments of the Mahalanobis distance (Mahalanobis, 1936) and the method of linear discriminant analysis (Huberty and Olejnik, 2006).

Linear discriminant analysis (LDA), developed by Fisher (1936), is still a commonly used multivariate statistical method for predicting group membership from numeric variables. It was first applied to the topic of ancestry estimation from craniometric variables by Giles and Elliot (1962). Their study provided a set of discriminant functions that other anthropologists could use to estimate American White, American Black, or American Indian affiliation of males and females. LDA is still widely used in biological anthropology to analyze and evaluate population differences and estimate population affinity. It is the method of prediction used by FORDISC (Ousley and Jantz, 2005), a computer program designed to estimate ancestry and other biological characteristics.

Though LDA has been useful in biological and forensic anthropology, it makes mathematical assumptions pertaining to data distributions that can invalidate results if violated. It assumes that the data for each group included in the analysis are

multivariate normally distributed and that the groups have equal covariance matrices. Other discriminant analysis techniques – such as k-nearest neighbor analysis (KNN), quadratic discriminant analysis (QDA), and classification trees (CT) – have since become an option, but remain mostly unevaluated in the context of forensic craniometric ancestry estimation. KNN is an exception that has been applied in past studies (e.g., Ousley et al., 2009). The above mentioned discriminant analyses have assumptions of their own that differ from those of LDA, which may justify their use in situations where LDA's assumptions are violated. For example, QDA assumes multivariate normality, but does not require equal covariances. Therefore, QDA may be a more appropriate method in situations where data have multivariate normal distributions, but unequal covariances. The importance of validating LDA's assumptions has been mentioned by Feldesman (2002). Unfortunately, the same article points out that few analyses using LDA test its assumptions or even recognize them to begin with. Feldesman (2002) states that many researchers simply proceed with an LDA assuming that the assumptions are not violated.

The purpose of this analysis is to evaluate the assumptions and performances of LDA, QDA, KNN, and CT in forensic ancestry estimation using craniometric variables. Quadratic discriminant analysis is a method that requires multivariate normal distributions, but does not assume homogeneity of variance. K-nearest neighbor analysis is a nonparametric method that does not assume multivariate normality, but does assume homogeneity of variance. Classification trees are discriminant methods that makes no distributional or variance assumptions. Data consisting of craniometric measurements on contemporary human groups from the Forensic Data Bank (Jantz

and Moore-Jansen, 1987) will be used. The performances, advantages, and disadvantage of the methods will be compared and discussed. Recommendations for future use and research will be made based upon these criteria.

CHAPTER I

LITERATURE REVIEW

Early Statistical Explorations of Human Variation

Giles and Elliot (1962) were the first anthropologists to use multivariate statistical analysis in an attempt to estimate race; their inquiry focused on the identification of American Blacks, American Whites, and American Indians. They believed the cranium provided the best indication of race. Thus, they conducted an LDA on 225 males and 225 females using eight cranial variables: glabello-occipital length; maximum cranial width; basion-bregma height; maximum bi-zygomatic diameter; prosthion-nasion height; basion-nasion; basion-prosthion; and nasal breadth. The analysis resulted in the generation of four discriminant functions, two for each sex, which predicted American White versus American Black and American White versus Native American. Giles and Elliot (1962) reported high classification rates (re-substitution – 86.7% males, 89.8% females; holdout – 79.8% males, 86.6% females), but subsequent tests of their functions failed to match the reported performance (Birkby, 1966; Snow *et al.*, 1979; Ayers *et al.*, 1990; Fisher and Gill, 1990) The functions' performance may not have been ideal, but the method of analysis was pioneering (Iscan, 1988), it set the stage for future studies of ancestry. Multivariate statistics, especially LDA, are still relied on for assessing human craniometric variation and ancestry estimation.

Almost a decade after Giles and Elliot (1962) W. W. Howells published the results of an extensive investigation of global human craniometric variation (Howells, 1973). This study sought to establish a basis for comparison of skulls by constructing a sample of total cranial variation and subjecting it to multivariate analyses. Howells's

main question was whether differences in cranial form among populations are based upon the same factors as differences within populations. His goals were to find variables that could differentiate populations in terms of cranial shape and to describe how shape differs across populations. His sample contained 1,652 crania of individuals from 17 populations distributed across five major geographic regions: Europe; Africa; Asia; the Pacific; and America. Howells collected data comprising 70 measurements and angles and used them in a variety of multivariate analyses, such as LDA and cluster analysis. Though Howells (1973) acknowledged that his sample missed Southeast Asia and large populations, such as China and India, his results still showed a geographic patterning to human cranial variation.

Howells published another influential study in 1989, where he explored whether sets of characteristics can distinguish a given population from others and whether there are traits specific to geographic regions. This time he specified six world regions, which he identified as: Europe; Africa (sub-Saharan); the Far East (Japan and China); Australo-Melanesia; Polynesia; and America. He sampled 18 populations, each region represented by three groups. If distinguishing characteristics were apparent, Howells wanted to know whether they reflected ancient separations of regional groups or recent microevolution. He conducted his analysis by subjecting the C-scores (describing cranial shape, while controlling for size) of 57 cranial measurements for each individual to cluster analysis and Q-mode analysis. His results revealed patterned regional variation in cranial shape, supporting his 1973 findings. However, Howells (1989) noted that the observed differences between populations were not great. He also noted that no individual population or regional group was particularly distant or distinguished from

the others. Aside from his results concerning patterned geographic variation, Howells (1973;1989) also contributed useful methodology and a great deal of data that is still used in contemporary discriminant analyses evaluating craniometric variation and population affinity (Sauer and Wankmiller, 2009).

Software for Craniometric Assessment of Ancestry

There are a couple software programs designed for craniometric ancestry estimation. CRANID, a computer program developed by Richard Wright (1992), uses Howells's data to classify an unknown cranium into a known population. The goal is to assess the geographic origin of the individual in question. CRANID was developed primarily for use in forensic cases where a decedent is unknown (Wright, 1992). It uses 29 variables to describe the shape of the cranium, then compares the shape to a reference sample of 2,870 crania from 66 groups (Wright, 2008). CRANID analyzes the crania using principal components analysis, cluster analysis, and k-nearest neighbor analysis, comparing an unknown cranium to its 50 nearest neighbors (Wright, 1992). Originally, Wright (1992) stated that the above methods performed better than LDA, referencing analyses that he chose not to report in his paper. However, Wright (2008) states that CRANID classifies a skull using LDA in a later publication.

Jantz and Ousley (1993) developed another discriminant analysis program they named FORDISC, now in it's third version. It uses LDA to classify an unknown skull into one of 11 groups. Anywhere from 1 to 34 craniometric variables can be used to classify an individual (Sauer and Wankmiller, 2009) and the unknown can be compared to a reduced number of groups rather than all 11. The reference dataset used by FORDISC contains samples from the Forensic Data Bank (FDB), a database of

measurements on individuals whose skeletons were analyzed by forensic anthropologists nationwide in casework or study (Jantz and Moore-Jansen, 1987). The Howells data set and samples from the Terry and Todd collections are also part of the reference data (Ousley and Jantz, 2005). The program is most commonly used by forensic anthropologists due to the FDB reference sample's more accurate representation of a contemporary American forensic population and the program's flexibility for which and how many variables and reference groups are used.

A Controversy in the Field: Anthropology and Forensic Race Estimation

Forensic anthropology is currently defined as, "the application of anthropological methods and theory - particularly those relating to the recovery and analysis of human remains - to resolve legal matters" (SWGANTH, 2012). The traditional role of forensic anthropologists is to apply methods of physical anthropology to the identification of decomposed or skeletonized human remains for medico-legal death investigations (Stewart, 1979). Identifying unknown remains is a significant part of forensic anthropological analyses, but the field has expanded to include other specialties, such as recovery, excavation, and skeletal trauma analysis (Grivas and Komar, 2008). The traditional role of the biological profile in identification is not as significant as it once was; other techniques, such as dental comparisons and individualizing pathological analysis, have been shown to more significantly improve the likelihood of positive identification (Steadman *et al.*, 2006). However, the biological profile, typically including estimates of age, sex, stature, and ancestry, is still a necessary tool because it narrows down the pool of unknowns (Sauer, 1992; Sauer and Wankmiller, 2009; Algee-Hewitt,

2011). While the estimation of the former three categories of the biological profile are not controversial among practitioners, ancestry estimation is wrought with controversy. Confusion and debate currently surround interpretation of craniometric variation.

It is generally accepted in the field that cranial shape and other morphometric characteristics of the human skeleton vary with geography. However, whether or not this variation can be used to accurately classify individuals into "racial" groups is a hotly debated topic. While there are anthropologists who argue that ancestry can be estimated from craniometric traits and is an important part of forensic anthropological analyses, there are others who argue that it cannot be estimated accurately, that it should not be estimated because it reifies racist ideas, or both.

The arguments against ancestry estimation range from philosophical discussions of its use to infer race to statistical analyses illustrating the difficulty of calculating accurate figures and the utility of the information in identification. Steadman *et al.* (2006) discuss the difficulty of obtaining accurate posterior probabilities of group membership due to the treatment of ancestry as a categorical variable that does not account for admixture. Furthermore, properly accounting for admixture makes defining the population at large (used to derive an accurate likelihood ratio) very difficult (Steadman *et al.*, 2006). Konigsberg *et al.* (2009) illustrate the calculation of an ancestry likelihood ratio in a case and how information on ancestry does not contribute significantly to identifications in terms of improving the likelihood of positive identification. Steadman *et al.* (2006) and Konigsberg *et al.* (2009) also discuss difficulties in obtaining prior and posterior probability estimates because of the need to use bureaucratic race and census data to derive informative priors. Smay and

Armelagos (2000) reference past genetic studies of human variation and discuss logical flaws that they see in arguments supporting the forensic estimation of race, while Goodman (1997) discusses the disconnect between human biological variation and race. Goodman also argues that race is a bad scientific variable stating, "generalizations ought not be based on an ill-defined, constantly changing and contextually loaded variable" (Goodman, 1997:23) and, "one cannot do predictive science based on a changing, indefinable cause" (Goodman, 1997:24). Finally, Williams *et al.* (2005) and Elliott and Collard (2009) have attempted to show that FORDISC performs poorly. However, Williams *et al.* (2005) have been heavily criticized for their statistical methodology (Ousley *et al.*, 2009) and Elliott and Collard (2009) have received some criticism for over-fitting (Algee-Hewitt, 2011) and misunderstanding statements made at a FORDISC workshop (Ousley and Jantz, 2012).

On the other hand, there are studies that suggest cranial shape does reflect geographic patterning and genetic variation to a degree. Two examples of such work are the studies done by Howells (1973;1989) discussed above. More recent examples include work conducted by Heather Smith and colleagues (Smith *et al.*, 2007; Smith, 2009) and Roseman and Weaver (2004). In their exploration of temporal bone morphology and variability Smith *et al.* (2007) find significant differences between modern human populations. A discriminant analysis on 40 principal components achieved an overall classification rate of 73% with individual population hit rates ranging from 56 to 85%. The authors conclude that the temporal bone has significant discriminatory power amongst modern populations. In her later study, Smith (2009) assessed the utility of the entire skull as well as its various regions in differentiating

modern human populations. To do so she compared morphological distances of the cranium with genetic distances by correlation analysis. The results show that the entire cranium, temporal bone, upper face, and basicranium are significantly correlated with molecular distances. The correlations are not strong, however, and the r^2 values are pretty small, indicating that these factors do not account for a large amount of variation in cranial morphology. Roseman and Weaver (2004) find that there are morphological features of the cranium that vary to a greater degree than is expected under selectively neutral conditions. Though their F_{st} estimates are similar to those derived in works showing that among population craniometric variation is similar to that of neutral genetic markers (Relethford, 1994;2002), they find that a number of individual traits do vary among populations to a greater degree. Traits that have greater between-group variance are from the upper nasal region (greatest variance), mastoid size, and frontal curvature. The authors (Roseman and Weaver, 2004) note that nasal variables are commonly used to estimate ancestry. However, they also observe a large degree of overlap in the density plots, indicating that individuals from different regions can be quite similar. As a final example, Spradley *et al.* (2008) report fairly good classification rates using a stepwise LDA on a sample consisting of American Whites and Southwest Hispanics, though they report difficulty classifying Southwest Hispanics when more reference populations are included.

The Biological Race Concept: Contemporary Views in Anthropology

The concept of human biological race, that subspecies exist within *Homo sapiens*, is currently accepted by very few physical anthropologists (Lieberman *et al.*, 2003). Race is widely regarded as a dynamic social construct rather than a stable

biological phenomenon (Goodman, 1997). Nonetheless, anthropologists use skeletal data to gain insight into human evolutionary trends, population histories, and population structure. There are, however, many anthropologists who frown upon the use of skeletal data in analyses of contemporary human groups, especially when they are culturally defined. Algee-Hewitt (2011:2) states:

> A perceived gap exists…in contemporary anthropological thinking between what can be conceptualized intuitively based upon social constructions of human diversity and ideas of the self and others, and what can be tested and corroborated empirically using best practices and following accepted standards of scientific rigor.

Therefore, a dispute exists in forensic anthropology as to whether anthropologists should estimate race as part of the biological profile.

While anthropologists agree that systematic, geographic human variation exists (Sauer, 1992), some contend that these differences can be used to estimate an individual's geographic origin (Brace, 1995; Ousley *et al.*, 2009). Others argue there is not enough among group variation to enable classification (Smay and Armelagos, 2000; Williams *et al.*, 2005). Those who accept the idea of estimable ancestry and use it to predict race in forensics argue that knowing an individual's ancestry allows anthropologists to make statements as to their probable racial classification in a social system (Sauer, 1992). As Ousley and Jantz (2012:311) explain, "Sex and ancestry in forensic anthropology are *estimated* because they are determined or defined by other means", such as the presence or absence of a Y chromosome in the case of sex.

Konigsberg *et al.* (2009) briefly touch on the controversy surrounding forensic race estimation.

The terms “race” and “ancestry” are often used interchangeably, and the latter may be viewed as a euphemism of the former, but Konigsberg *et al.* (2009) describe them as two different concepts. Race is a socially constructed system of classification and self-identification, while ancestry has a biological basis, referring to observable variation in natural biological characteristics of the human form (Konigsberg *et al.*, 2009). Anthropologists, therefore, can only estimate ancestry. This is done by “using morphological signatures of population history that represent a complex accumulation of genetic variation shaped by generations of microevolution and environmental pressures” (Konigsberg *et al.*, 2009:78).

However, what is ultimately being reported to law enforcement is a social race classification of White or Black, for example. There are biological signatures that are unique to various human populations throughout the world. However, in forensic ancestry estimation, practitioners are evaluating a far more limited range of variation, typically unique to one nation. Likewise, the predicted groups are those of that nation’s classificatory system. While the social races in the United States were constructed and are primarily defined by skin color, the populations originated from divergent regions of the world. Thus, they would likely have unique variations, morphological signatures, and historical patterns of microevolution. Many of these signatures may be preserved in modern American populations due to selective mating. However, these signatures may become far more difficult to recognize as intermixing increases. An example of this can be seen with the difficulties in identifying individuals of Hispanic ancestry caused by the

broad range of variation and genetic and population history represented amongst the individuals to which the term is currently ascribed.

FORDISC Methodology and Potential Limitations

FORDISC uses discriminant analysis for ancestry estimation (Ousley and Jantz, 2005;2012). Discriminant analysis describes a category of statistical methods that assess the degree to which it is possible to separate two or more groups of observations given a set of variables (Rencher, 2002; Manly, 2005; Huberty and Olejnik, 2006). The specific analysis FORDISC performs is LDA. While LDA may be a very popular discriminant method (Ousley and Jantz, 2005;2012), it makes a number of assumptions about the data that must be met for the results to be valid and interpretable.

LDA is a parametric method that assumes that the data for each group have a multivariate normal distribution. It also assumes homogeneity of variance, that the covariance matrices for all groups in the reference sample are equal (Rencher, 2002; Manly, 2005; Huberty and Olejnik, 2006). While these two phenomena can be difficult to test, a researcher or practitioner using LDA must evaluate the assumptions (Ousley and Jantz, 2012). FORDISC 3.0 added a test for homogeneity of variance (Ousley and Jantz, 2005). However, it assumses that multivariate normality holds for each of the groups in the FORDISC reference sample (Ousley and Jantz, 2012). Nonetheless, Ousley and Jantz (2012:324) state, "*as long as the requirements are met* and other statistical results are acceptable, Fordisc will always indicate the most similar group based on the groups and measures used [emphasis added]," and "most importantly, no matter how good the apparent correct classification rate is, if certain rules are broken,

there are reasons to doubt such performance will hold when applied to the individual being classified" (Ousley and Jantz, 2012:325). Thus, while not satisfying the assumptions does not necessarily preclude the LDA from yielding good estimates of classification accuracy, it calls into question the reported statistical results and out-of-sample performance of the classification rules.

There are a variety of alternatives to LDA. Some researchers have utilized discriminant methods other than LDA, such as k-nearest neighbor analysis (Wright, 1992; Ousley *et al.*, 2009) and finite mixture analysis (Konigsberg *et al.*, 2009), though an explicit evaluation and comparison of which discriminant techniques are useful in various situations has yet to be performed. Feldesman (2002), however, conducted a comparative analysis of LDA and CT using 10 measurements from the distal humerus on five groups of modern hominoids. While LDA and CT yielded very similar classification rates, CT was recommended for use when LDA assumptions were violated or there were missing values in the data.

The Need for Accurate and Reliable Statistical Figures in Forensics

The Supreme Court rulings of *Daubert v. Merrill Dow Pharmaceuticals, Inc.* (1993) and *Kumho Tire, Ltd v.Carmichael* (1999) have significantly affected expert testimony and evidence admissibility in trials (Christensen, 2004; Grivas and Komar, 2008). *Daubert* has received the most attention within forensic anthropology (Grivas and Komar, 2008) because it directly concerns scientific witness testimony. It provides five guidelines by which testimony may be evaluated. *Daubert* states that the content of testimony must: (1) be testable and have been tested through the scientific method; (2) have been subjected to the peer review process; (3) have established standards; (4)

have a known or potential error rate; and (5) have widespread acceptance by the relevant scientific community (Christensen, 2004; Grivas and Komar, 2008). Thus, the application of valid and interpretable statistics is necessary for forensic anthropologists to adhere to the *Daubert* guidelines.

CHAPTER II

MATERIALS AND METHODS

Materials

The data are comprised of White, Black, and Hispanic males and females from the Forensic Data Bank (FDB) with observations on the 24 original caraniometric variables used in FORDISC (Ousley and Jantz, 2005) and outlined in "Data Collection Procedures for Forensic Skeletal Matrial" (Moore-Jansen *et al.*, 1994). The dataset has a total of 3020 observations consisting of 1925 males, 1093 females, and 2 individuals missing observations for sex (all of the available FDB data for American Whites, American Blacks, and Hispanics at the time). There are 1939 American Whites (1200 males, 739 females), 610 American Blacks (362 males, 248 females), and 469 Hispanics (363 males, 106 females). Finally, there are 2283 positively identified individuals, 183 presumptive identifications, 499 unidentified individuals, and 55 observations missing values for ID Status. Tables 1 to 3 display this information.

Table 1. Sex, ID Status, and Race samples and missing values.

Statistics

		Sex	IDStatus	Race
N	Valid	3018	2965	3020
	Missing	2	55	0

Table 2. Group samples.

Sex * Race Crosstabulation

Count

		Race			Total
		White	Black	Hispanic	
Sex	Male	1200	362	363	1925
	Female	739	248	106	1093
Total		1939	610	469	3018

Table 3. ID Statuses by group.

IDStatus * Race Crosstabulation

Count

		Race			Total
		White	Black	Hispanic	
IDStatus	Positive	1644	458	181	2283
	Presumptive	102	56	25	183
	Unidentified	163	78	258	499
Total		1909	592	464	2965

This analysis is limited to positively identified individuals with known sex, race, and birth year. To control for secular change (Jantz and Jantz, 2000; Jantz, 2001), only individuals born after 1930 will be used. Though there is no formal analysis of an appropriate cutoff to ensure contemporaneity, FORDISC (Ousley and Jantz, 2005) uses this cutoff and other analyses have used similar cutoff points (Spradley *et al.*, 2008). Of the positively identified observations, 232 are missing values for birth year and will be excluded. Applying these criteria leaves 1480 observations. The group sample sizes are displayed in Table 4.

Table 4. Group samples using only positive IDs.

Sex * Race Crosstabulation

Count

		Race			Total
		White	Black	Hispanic	
Sex	Male	715	143	88	946
	Female	407	94	33	534
Total		1122	237	121	1480

Missing Value Analysis and Imputation

A missing value analysis in SPSS 20 (Corp., 2011) shows that the percent of missing values on the variables ranges from 3.5-34.5% (Table 5). A missing value matrix was created with binary variables indicating whether a value is present (0) or absent (1) to help observe missingness on observations. The values for these variables were summed across rows to yield a total number of missing values on each observation. The sum was divided by the total number of variables to give a percentage of observed values for each observation. Analyzing missingness by observation for all 24 variables with a histogram (Figure 1) shows that about 68% of observations are missing 5% or fewer values. About 22% of observations are missing between 5% and 21%, while the rest range from 21% to 96% of values missing.

The presence of missing values in a data set can affect the results of a statistical analysis. In the case of discriminant analyses, classifier accuracy on out of sample observations can be significantly reduced (Twedt and Gill, 1992; Acuña and Rodriguez, 2004). Whenever there are missing values in a data set, the analyst must come up with a strategy for dealing with them that depends on the pattern, mechanism, and degree of missingness. Acuña and Rodriguez (2004) describe three general classes of methods for handling missing values: (1) case deletion; (2) parameter estimation; and (3) imputation techniques. Various methods fall under each category, such as complete-case analysis for the first, expectation maximization for the second, and regression estimation for the third. They also describe the impact that degrees of missingness can have on selecting an imputation method stating, “Rates of less than 1% missing data are generally considered trivial, 1-5% manageable. However, 5-15%

require sophisticated methods to handle, and more than 15% may severely impact any kind of interpretation" (Acuña and Rodriguez, 2004:639). The problem of missing values in this analysis is handled with a combination of case deletion and regression imputation techniques.

Table 5. Missing values by variable.

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
GOL	1428	183.17	9.008	52	3.5	7	3
XCB	1419	137.73	6.230	61	4.1	9	17
ZYB	1367	126.56	7.181	113	7.6	20	1
BBH	1414	137.84	6.456	66	4.5	6	1
BNL	1408	102.66	5.861	72	4.9	10	1
BPL	1256	95.93	7.221	224	15.1	4	2
MAB	970	61.24	5.108	510	34.5	5	3
MAL	1185	52.99	4.751	295	19.9	13	12
AUB	1332	120.70	6.085	148	10.0	16	6
UFHT	1175	69.79	7.380	305	20.6	18	4
WFB	1387	95.19	7.476	93	6.3	8	7
UFBR	1201	103.17	5.183	279	18.9	12	18
NLH	1359	51.17	3.813	121	8.2	21	1
NLB	1390	23.86	2.352	90	6.1	1	11
OBB	1368	40.20	2.565	112	7.6	32	8
OBH	1377	34.09	2.255	103	7.0	6	5
EKB	1302	96.41	4.743	178	12.0	9	3
DKB	1324	21.04	2.747	156	10.5	4	9
FRC	1374	112.34	5.941	106	7.2	3	6
PAC	1372	115.45	7.119	108	7.3	10	12
OCC	1364	99.22	5.959	116	7.8	4	14
FOL	1364	36.65	2.663	116	7.8	20	22
FOB	1271	31.07	2.573	209	14.1	2	4
MDH	1356	30.10	4.891	124	8.4	12	8

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

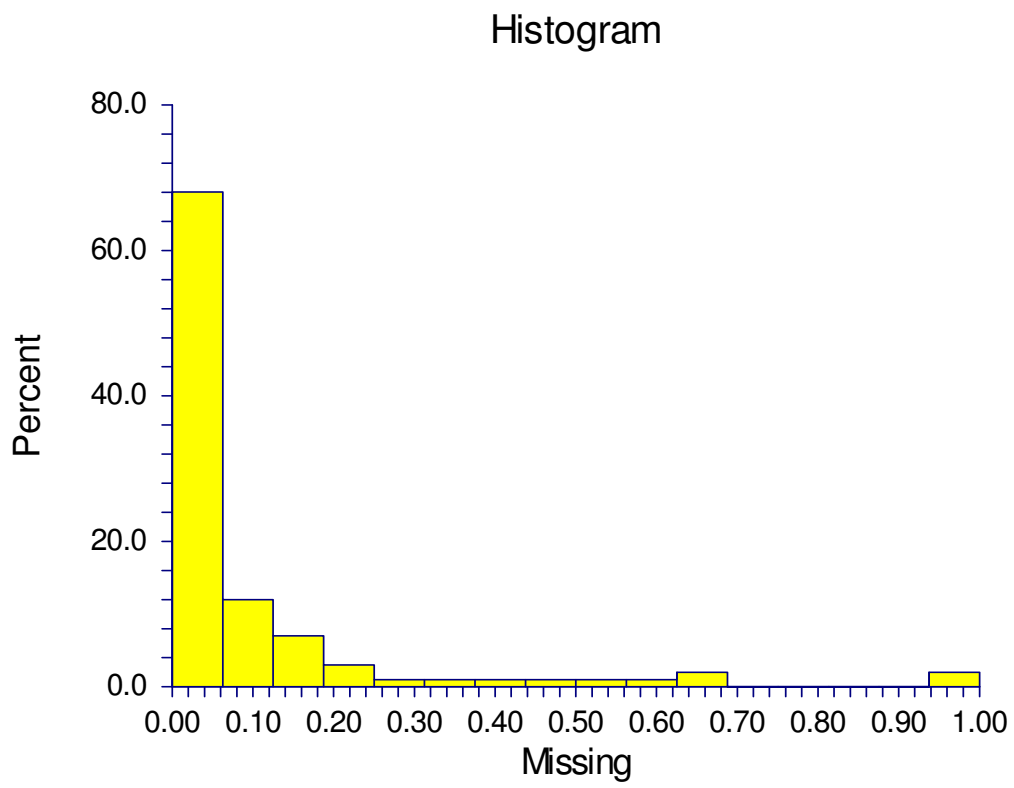

Figure 1. Percent of observations by percent missing values

Case Deletion

Case deletion techniques are easy to use and widely applied (Acuña and Rodriguez, 2004). They tend to limit the analysis to include only cases with observed values on all variables (complete-case analysis) or, in the case of variable deletion, variables with observed values on all cases (complete-variable analysis). Rather than deleting all cases or variables with missing observations, the analyst may choose to identify and delete only those with large degrees of missingness (Acuña and Rodriguez, 2004).

Case deletion is simple because there is no need to estimate data, and it facilitates comparison because the statistics are calculated on a common sample (Little

and Ruban, 2002). However, if there is a large degree of missingness in the data or the sample size is small, then case deletion can cause large amounts of data loss by significantly reducing the sample size and deleting cases with observed values for other variables (Twedt and Gill, 1992; Little and Ruban, 2002; Acuña and Rodriguez, 2004; Templ *et al.*, 2011). The information loss can reduce the accuracy of the analysis and introduce bias when the missingness mechanism is not missing completely at random (MCAR) (Little and Ruban, 2002).

A complete-case analysis would reduce this sample to 542 observations (down from 1480), significantly reducing the overall sample size and making some group samples unacceptably small. Therefore, variables and observations with excessive degrees of missingness were identified and deleted in the hopes that far fewer observation deletions are required. Observations with large degrees of missingness cannot have their missing values reliably estimated and have to be deleted.

Which Variables and Observations to Retain?

The first step taken to address the missingness was to find a key subset of variables that past researchers have found useful in ancestry estimation. Hopefully, reducing the data set to the key subset removes some of the variables with high degrees of missingness, reducing the missingness on observations, and allowing for the retention of more observations. By observing missingness on a key subset, the deletion of observations is evaluated on important variables, reducing the likelihood that important information is lost due to missingness on potentially uninformative variables.

Ousley et al. (2009) report that a stepwise discriminant analysis conducted on American Whites and American Blacks resulted in a 95% hit rate using 6 variables:

cranial base length (BNL); basion-prosthion length (BPL); biauricular breadth (AUB); nasal breadth (NLB); maxillo-alveolar breadth (MAB); and orbital height (OBH). However, their analysis only included American Blacks and American Whites. Spradley *et al.* (2008) conducted an analysis on American Blacks, American Whites, Hispanics and Guatemalans. Their significant variables may help predict Hispanic ancestry. They found maximum cranial length (GOL), orbital breadth (OBB), bizygomatic breadth (ZYB), biorbital breadth (EKB), maximum cranial breadth (XCB), and interorbital breadth (DKB) significant in predicting group membership. These were in addition to the above variables reported by Ousley *et al.* (2009). Some other variables not used by FORDISC were also included. Spradley *et al.* (2008) used a stepwise discriminant analysis, but report lower classification rates than Ousley *et al.* (2009) and weak posterior probabilities.

Figure 2 and Table 6 show that the key subset of 12 variables has greater rates of missingness than the overall data set. Only about 52% of observations have 6% or less missingness, about 30% of observations are missing between 6%-25% of their values, and the rest missing up to 100%. Limiting the analysis to these variables does not mitigate the problem of missing values by observation. Therefore, a strategy employing a combination of variable and case deletion based upon degree of missingness will be used to reduce the missingness to a more manageable level.

Table 6. Missingness for the key subset.

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
GOL	1459	183.17	9.022	52	3.4	7	3
XCB	1450	137.73	6.230	61	4.0	9	17
ZYB	1397	126.53	7.159	114	7.5	20	1
BNL	1439	102.62	5.872	72	4.8	10	1
BPL	1280	95.89	7.223	231	15.3	4	9
MAB	983	61.21	5.103	528	34.9	5	3
AUB	1363	120.68	6.084	148	9.8	17	6
NLB	1418	23.87	2.347	93	6.2	1	11
OBB	1395	40.19	2.554	116	7.7	32	8
OBH	1405	34.09	2.250	106	7.0	6	5
EKB	1331	96.42	4.722	180	11.9	9	3
DKB	1353	21.05	2.746	158	10.5	4	9

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

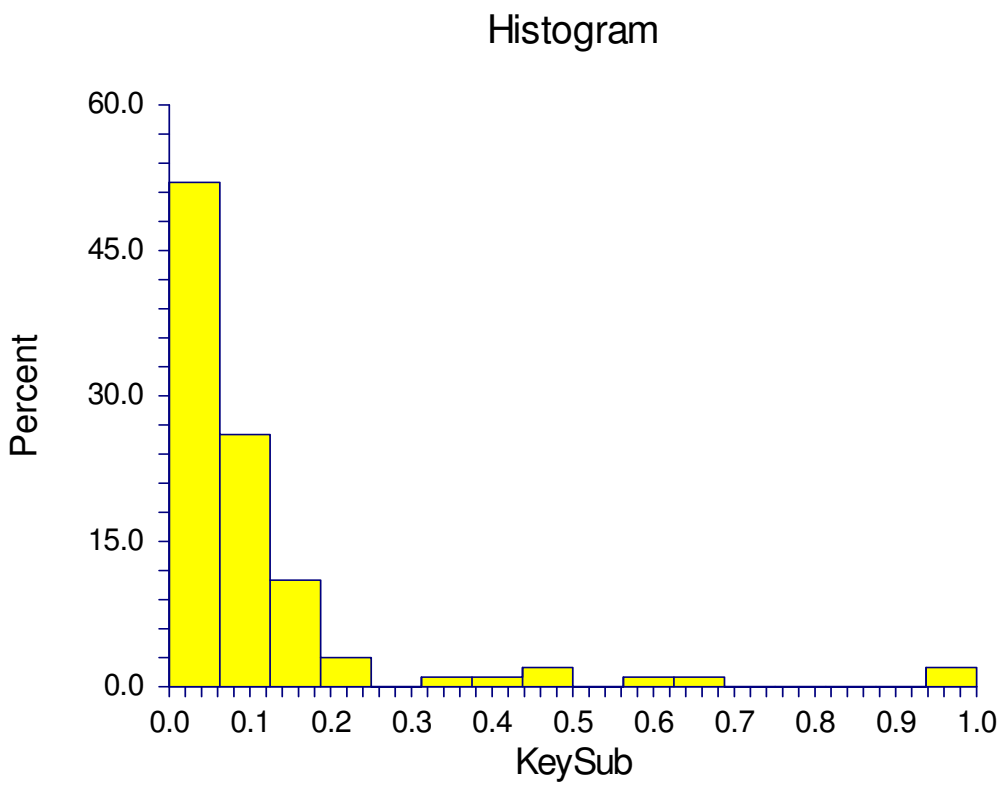

Figure 2. Percent of observations by percent of missing values: 12-variable subset.

First, only variables with 10 percent or fewer values missing were retained to limit the degree of missingness and allow for the application of practical estimation techniques. There are two more key variables in the 10-15% range, EKB and DKB. In an attempt to retain these measurements missingness was evaluated in a subset including them. This gave a 16-variable subset with a maximum of 10% missingness including: GOL; XCB; ZYB; AUB; NLB; OBB; OBH; BNL; Basion-Bregma Height (BBH); Minimum Frontal Breadth (WFB); Nasal Height (NLH); Frontal Chord (FRC); Parietal Chord (PAC); Occipital Chord (OCC); Foramen Magnum Length (FOL); and Mastoid Length (MDH). Adding EKB and DKB gives an 18-variable subset with a maximum of 12% missingness.

Two McNemar tests were conducted to see if there was a significant difference between the numbers of observations included in each variable subset on various degrees of missingness up to 20% at 5% intervals. The results were used to decide which variable subset to retain and the missingness threshold for observation retention. The first test included observations with up to 5% and 10% missingness. The second evaluated observations with up to 15% and 20% missingness. The first test was significant, but the second test was not (results in Table 7). Therefore, there is a dependence structure between the number of variables and the degree of missingness for lower degrees of missingness (i.e. 5% and 10%), but not high degrees of missingness (i.e. 15% and 20%). Thus, the 16-variable subset and observations with up to 10% missingness on those variables were chosen.

If the 18-variable subset was used, obseravations with 15% or 20% missingness were retained, or both, then missingness would be more pervasive and problematic. If observations with only 5% missingness were retained, the sample size would be significantly reduced, as indicated by the significant McNemar result. Thus, the 16-variable subset reduces the degree of missingness while allowing more observations to be retained.

Table 7 - McNemar results.

Subset	**% Missing**		**Statistics**	
	0.05	**0.10**	X^2	***p*-value**
16-Variable	1161	1258	7.423	0.006
18-Variable	1125	1227		

Subset	% Missing		Statistics	
	0.15	0.20	X^2	*p*-value
16-Variable	1312	1335	0.954	0.329
18-Variable	1285	1316		

Using this subset of variables and observations, the total sample size is reduced to1258. Group sample sizes are shown in Table 8. The remaining missing values were imputed with iterative regression techniques. However, the Hispanic female sample will not be included in the analysis due to its small sample size (n = 27). This falls below the recommended sample size of three to five times the number of variables for a discriminant analysis (Ousley and Jantz, 2005; Huberty and Olejnik, 2006; Ousley and Jantz, 2012).

Table 8. Final Sample.

Sex * Race Crosstabulation

Count

		Race			Total
		White	Black	Hispanic	
Sex	Male	629	117	79	825
	Female	331	75	27	433
Total		960	192	106	1258

Regression Imputation

In regression imputation procedures the variable with a missing value is used as the response and those with observed values in the concordant row are used as regressors (Acuña and Rodriguez, 2004; Hintze, 2007; Templ *et al.*, 2011). The process may go through multiple iterations to achieve greater accuracy. The imputed

value from one iteration is used to initialize the next until convergence is reached (Hintze, 2007; Templ *et al.*, 2011). These methods can be very accurate, but outliers and violations of normality can negatively affect the estimations (Hintze, 2007; Templ *et al.*, 2011). Though a cqplot macro in SAS 9.3 showed that each of the groups failed to meet the multivariate normal assumptions, two iterative regression procedures were applied to the data to estimate the missing values: (1) the NCSS (Hintze, 2007) multivariate normal missing value estimation procedure in data screening; and (2) the `irmi()` function of the VIM package version 3.0.1 (Templ *et al.*, 2012) in R version 2.15.1 (R Core Team, 2012).

Two similar forms of estimation were chosen so that they could be used to validate one another by comparing estimates. Furthermore, there are outliers in the data, which NCSS's procedure does not take any steps to address. On the other hand, `irmi()`, can use robust methods to reduce the influence of outliers, which can affect normality and regression coefficient estimations (Templ *et al.*, 2011).

The documentation provided for the NCSS (Hintze, 2007) iterative regression estimation procedure is limited. The help documentation states, "a regression analysis is conducted using the variable containing the missing values as the dependent variable and all variables with nonmissing data in this row as independent variables" (Hintze, 2007:118-2). The observation's observed values are entered into the regression equation to calculate an estimate for the missing value. This method is repeated over multiple iterations, using the predicted value from one run for the estimations of the next. McCarty (2005) found this procedure quick and accurate.

IRMI stands for Iterative Robust Model-based Imputation (Templ *et al.*, 2011). VIM's (Templ *et al.*, 2012) `irmi()` works similarly to the NCSS procedure, but is robust to outliers. It initializes the imputation with either *k*-nearest neighbor or median estimates for the missing values. An iterative robust regression procedure is then applied to the data to estimate the missing values. The function accounts for the distribution of the response variable, using robust regression if it is continuous and other methods, such as robust logistic regression, for other types of variables. `Irmi()` was used because data screening in NCSS (Hintze, 2007) indicated the presence of multivariate outliers in a few of the groups. The outlier diagnostic used is a T^2 statistic based upon observations' Mahalanobis distance from the centroid (Hintze, 2007). Outliers were evaluated within groups, not on the overall sample.

Estimations for this Data

The above imputation procedures were applied to the data within groups to maintain the intergroup relationships, not reducing the between-group variation. Each procedure went through 5 iterations. McCarty (2005) found no significant improvements in accuracy past 5 iterations in NCSS. Templ *et al.* (2011) state that experiments with their IRMI algorithm show that convergence is usually achieved in a few iterations, and 5 is the default number of iterations in `irmi()`. Significant accuracy gains were not achieved after the second iteration in the test conducted in their 2011 paper (Templ *et al.*, 2011). The IRMI estimations were initialized with median estimates.

After the imputations were completed, descriptive statistics and outliers were checked on the imputed data sets and compared with one another and those of the original data to make sure reasonable estimates were made and the variables for which

values were imputed were not drastically changed. The means, standard deviations, ranges, and distributions were consistent across datasets. However, some groups had more outliers after the estimations. For example, the White male group had 5 more outliers in each of its imputed data sets than in the original data. The 5 new outliers could not be evaluated in the original data set due to missingness because the multivariate outlier diagnostic cannot evaluate observations with missing values. However, outlying observations were consistent across all data sets and the new outliers are consistently identified as such in both imputed data sets.

The remainder of the analysis will be conducted on the robust imputation data set. Though both the robust regression and regular regression estimation techniques yielded similar results, it is necessary to be wary of the outliers.

Another Considered Imputation Option

K-nearest neighbor imputation was considered for use in missing value estimation. It is non-parametric, having no distributional assumptions. Thus, it would not risk being affected by the lack of multivariate normality in these samples. The procedure estimates a missing value on an observation, *i*, based upon a specified number, *k*, of nearest observations with observed values on the variable where *i* is missing a value. The nearest neighbor observations are determined by a distance measure and the mean of their observed values on the variable for which *i* is missing replaces *i*'s missing value (Troyanskaya *et al.*, 2001; Acuña and Rodriguez, 2004; Hron *et al.*, 2010). The distance is calculated using only the variables for which *i* has observed values. The estimation may also use a weighted mean, where the weights

are based upon distance, or the median to reduce the influence of outliers or more dissimilar observations.

KNN imputation does not directly consider the correlation structure of the data in making its estimations (Hron *et al.*, 2010) and it does not derive predictive models (Acuña and Rodriguez, 2004). This can be an advantage in situations where parametric assumptions are violated or many predictive models would be needed. However, in this case, the nature of the imputation process itself makes it unusable; KNN imputation uses inter-observational relationships in exchange for direct consideration of multivariate relationships, assuming that observations near to one another distance-wise are similar in form. This is an unsubstantiated assumption in this case and a recognized disadvantage of the method. It is possible for nearest neighbors to include information that is worse for estimation than some further neighbors (Troyanskaya *et al.*, 2001).

Methods

Four techniques were evaluated for their performance predicting group membership from craniometric variables: linear discriminant analysis; quadratic discriminant analysis; k-nearest neighbor analysis; and classification trees. These four methods fall under the general category of discriminant analysis. Manly (2005:105) states, “the problem that is addressed with discriminant analysis is the extent to which it is possible to separate two or more groups of individuals, given measurements for these individuals on several variables.” Depending on the research goals, discriminant analysis can be performed to explore differences among groups of observations or to establish rules for predicting group membership (Rencher, 2002; Huberty and Olejnik,

2006). This research is of the latter goal and is primarily concerned with classification accuracy. The linear, quadratic, and k-nearest neighbor analyses were performed in SAS Enterprise Guide 5.1, while the classification tree analysis was performed in SAS Enterprise Miner 12.1.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a popular discriminant method that has been applied to the question of population prediction with craniometric variables since 1962 (Giles and Elliot, 1962). It is still used in anthropology and is the method of analysis employed by FORDISC (Ousley and Jantz, 2005;2012). LDA creates linear discriminant functions (LDFs) that are used to classify observations with the set of predictor variables. The LDFs are combinations of variables that best separate groups (Rencher, 2002). Their parameters are optimized to maximize between-group variation, maximizing the distance between the groups. The LDFs generate scores for each observation in the data set that are used to predict group membership. The first LDF accounts for the greatest percentage of variation with the following functions accounting for less variation sequentially, the last function accounting for the least. The number of variables or groups in the analysis limits the potential number of LDFs. There cannot be more LDFs than variables or k-1 groups, where k is the total number of groups in the analysis (Manly, 2005).

The LDA procedure makes two mathematical assumptions in the derivation of the LDFs and its tests of significance: (1) that the sample covariance matrices for all groups are equal, a condition referred to as homogeneity of variance; and (2) that the data for each group are multivariate normally distributed (Rencher, 2002; Manly, 2005;

Huberty and Olejnik, 2006). While LDA may still classify well when its assumptions are violated (Manly, 2005), these conditions must be met for LDA to perform optimally (Huberty and Olejnik, 2006). Furthermore, if the assumptions are violated then the *p*-values may be over or under estimated and interpretability of the results is lost (Manly, 2005; Huberty and Olejnik, 2006).

Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is another discriminant analysis technique that, like LDA, constructs classification functions with the variables. However, in QDA the functions are not linear. In some cases linear functions may not provide the best group separation and quadratic functions are necessary. However, the factors typically influencing the selection of QDA are assumptions; while QDA does assume multivariate normality, it does not assume homogeneity of variance (Huberty and Olejnik, 2006). Therefore, QDA is a more appropriate analysis under conditions of variance heterogeneity. The ratio of sample size to the number of predictors used can also drive a decision as to whether to use LDA or QDA. Huberty and Olejnik (2006) state that if the ratio of observations to variables is small, then LDA may be preferable, even when variance heterogeneity is present. However, if the ratio is large and heterogeneity is present, then QDA is preferred. This guideline is obscure as they also state, "very little guidance as to definitions of 'small' and 'large' is proffered" (Huberty and Olejnik, 2006:281).

In his dissertation, Mark Connally (2004) reports that QDA outperforms LDA in every situation he evaluated under conditions of variance heterogeneity. However, he evaluated classification accuracy with an internal measure. Internal measures of

classification accuracy, such as re-substitution, are optimistically biased and are inappropriate for assessing out-of-sample classification rates (Rencher, 2002; Manly, 2005; Huberty and Olejnik, 2006; Fielding, 2007). While QDA may provide improved performance when the data display variance heterogeneity, there are a couple of reasons researchers are wary of the method: (1) it is more greatly affected by misclassifications and errors in the training data (Holden and Kelley, 2010); and (2) it is known for over-fitting (Ousley and Jantz, 2012).

K-Nearest Neighbor Analysis (KNN)

KNN was the first non-parametric discriminant analysis developed (Rencher, 2002). It does not assume multivariate normality, but it does assume homogeneity of variance. KNN differs from the above methods in that it does not find functions to discriminate groups, but classifies an observation based upon the group membership of a number, *k*, of its nearest neighbors. The distance between two observations is calculated as (Khattree and Naik, 2000):

$$d^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)\boldsymbol{V}^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are observation vectors and $\mathbf{V}$ is the covariance matrix. KNN posterior probabilities of group membership are calculated in SAS by the formula (Khattree and Naik, 2000):

$$p_i = \frac{\pi_i (k_i / n_i)}{\sum_{i=1}^{g} \pi_i (k_i / n_i)}$$

where π_i is the prior probability of belonging to group *i*, k_i is the number of observations amongst the k nearest neighbors belonging to group *i*, and n_i is the total number of

observations in group *i*. SAS classifies the observation into the group to which it has the greatest posterior probability of belonging. In the case where k=1 an observation would be classified into the group that its single closest neighboring observation belongs to. When k>1 an observation in question is classified into the group that is most common among its *k* nearest neighbors (Rencher, 2002; Huberty and Olejnik, 2006). Thus, KNN analysis assumes that observations from the same group are close to one another in multivariate space.

The researcher chooses the value of *k*. If *k* is too large then observations that are more distant from the one being classified, and theoretically less similar, will be considered. Researchers should seek to keep *k* small so that very close observations are used for the classification (Huberty and Olejnik, 2006). There is not an optimal value that can be generalized across analyses. If a researcher is unsure about which value of *k* to use, they may test multiple values and select the one that classifies the best (Rencher, 2002; Huberty and Olejnik, 2006). In this analysis k=3 was used. Results for k=1 were also assessed because of the strong performance reported by Ousley *et al.* (2009).

Classification Trees

Classification trees are non-parametric procedures that classify observations by repeatedly partitioning the data into subsets through a series of decisions (Breiman *et al.*, 1984). The goal is to create final subsets that are homogenous with respect to the group or class variable. Classification trees start with all of the observations grouped together in one node, referred to as the root node. All of the predictor variables are evaluated to determine which can be used to split the root node into groups that best

separate the classes. The best split is determined by evaluating a measure of impurity that quantifies the class makeup of each node. The impurity measure is maximized when all of the classes are equally mixed in a node and minimized when a node contains only one class (Breiman *et al.*, 1984). The chosen split is the one that most greatly reduces the impurity.

To choose a split rule the tree algorithm evaluates each variable one-by-one. It determines a value on each variable that would provide the best split. Next, it compares the splits for each variable to determine which does best and selects that variable (Breiman *et al.*, 1984).

A classification tree can construct binary or multinomial splits. This analysis uses binary splits. Therefore, a left and a right node are created each time a node is split. The new nodes' impurities are evaluated as well as another set of split decisions for each node. If the node's impurity can be reduced, then the split process is repeated. However, if the impurity cannot be significantly improved, then the splitting is stopped and the node becomes a terminal node. The terminal node is assigned a class equal to that of the class with the highest proportion of observations in the node. All of the observations in a terminal node are predicted as belonging to the class associated with that node. A new observation is classified by subjecting it to the decisions, or split rules, of the tree until the observation reaches a terminal node.

Breiman *et al.* (1984) discuss a number of advantages to using classification trees. Classification trees do not require multivariate normality and can be applied to any data structure. This makes them more generally applicable and removes the need to evaluate the distribution of the data for each group in the analysis. The procedure

does not require homogeneity of variance. Its conditional decisions made upon individual nodes do not require equal or similar data distributions. Classification trees inherently perform variable selection in their evaluation and selection of optimal splitting decisions. They are very robust to outliers and are even robust to misclassifications of observations in the training data. They can perform analyses in the presence of missing data and have a number of built-in procedures for handling missing values (Breiman *et al.*, 1984; Ding and Simonoff, 2010). Finally, the procedure is easy to use and provides output that can be easily interpreted and understood.

On the other hand, classification trees have a number of difficulties and drawbacks. First, the researcher must pay attention to the growth of the tree and make sure that it is not overly complex, or over-fit. Breiman *et al.* (1984) recommend growing the tree to its maximum and then selectively pruning it, recombining some of the later nodes. Though some programs, such as SAS Enterprise Miner, have functions built in that evaluate tree size and performance based on cross-validation or a holdout sample, this issue still requires attention when constructing a decision tree. Second, there are a variety of rules that can be selected for evaluating splitting. There are various measures of impurity and other methods, such as misclassification rates, that can be used to determine splits and the researcher must decide which is most appropriate. Third, for groups that can be separated by a linear function or some other combination of variables, it may take a very complex decision tree with many splits in order to achieve the same performance. In this case it is preferable to use another method, such as LDA. Lastly, divergent sample sizes can affect decision tree performance. Potential decisions are evaluated based upon how much they reduce the impurity,

which is calculated based upon the proportion of observations representing each group in a node. Tree splits may also be evaluated based upon misclassification rates. In either case, if groups in the analysis have very different sample sizes, simply predicting the largest group(s) and neglecting the smaller ones may provide the best performance in terms of classification accuracy. Improving classification accuracy for the smaller groups would only marginally improve the impurity measure or classification rate, so the algorithm may not classify them. Furthermore, when disparate sample sizes are present with groups that are difficult to separate, terminal nodes may often be assigned to the larger groups because they will likely have the greatest proportion of observations in the node due to their large size.

This classification tree analysis constructed a tree with binary splits. The Gini Index was the measure of impurity used for evaluating splits. A tree was constructed using the training and holdout samples. However, this tree, while having a strong overall classification rate, was not ideal because it did very poorly in predicting groups other than White males and White females. These two groups account for about 70-75% of the total sample. Therefore, a second tree was constructed on a stratified random sample selected 75 observations from each group. The number 75 was chosen because it is the sample size of the smallest group in the analysis (Black females). This latter tree was evaluated using v-fold crossvalidation. This procedure is like the jackknife procedure explained below, but rather than holding out one observation, it holds out a subsample of observations. The data set was split into 10 groups and a tree is constructed on 9 of the groups and evaluated on the one excluded group. This process is repeated until each of the 10 groups is used as the validation group. An

average performance across all ten runs is reported. Breiman *et al.* (1984) state that this method performs well with classification trees and it is the cross-validation method built into SAS Enterprise Miner 12.1.

Variable Selection

When running a discriminant analysis on a large set of variables it may be desirable to determine which variables are significant predictors of group membership. Including too many variables in a discriminant analysis can cause the classification accuracy to fall if there are redundant variables or others that do not contribute to group separation. The model can also become over-fit, yielding a strong re-substitution rate that is very positively biased (Huberty and Olejnik, 2006). An over-fit model is one that is fit to the training data so well that it is not generalizable to out-of-sample observations and performs poorly when classifying them.

A significant advantage of LDA is that with this method a researcher may select variables statistically, making variable selection much simpler. There are a number of ways that this may be done, but those incorporated in popular software packages are typically limited to forward, backward, and stepwise selection. These add variables, delete variables, or do both one at a time by evaluating significance in maximizing or significantly contributing to the ratio of between-to-within group variation (Rencher, 2002). These methods are only available for LDA because they are based on Wilk's Λ, which assumes multivariate normality and homogeneity of variance (Huberty and Olejnik, 2006). Therefore, in situations where QDA or KNN are being used, these assumptions are not likely met. Both forward and stepwise selection are built into FORDISC (Ousley and Jantz, 2005), but Ousley and Jantz (2012) suggest using

stepwise if variable reduction is desirable. Stepwise selection was used in the LDA for this analysis to find a significant subset of variables. The required *p*-value to enter a variable into the analysis was set to 0.2. Variables were only kept if they maintained a *p*-value less than 0.15.

Huberty and Olejnik (2006) do not advocate the use of stepwise selection, but they do recognize advantages of variable deletion. They attribute the widespread use of stepwise selection to its availability in two popular statistical software packages, SAS and SPSS. They state, however, "for a fixed total sample size, fewer outcome variables will lead to more precise estimates" (Huberty and Olejnik, 2006:104). They also recognize the advantages of parsimonious models that offer simpler interpretation and description and more accurate classification. Though it does offer a couple of significant advantages, variable selection should be used with caution. Manly (2005:114) points out that it has the potential to introduce bias and still cause over-fitting stating, "given enough variables, it is almost certain that some combination of them will produce significant discriminant functions by chance alone."

There are no built-in methods for variable selection with KNN or QDA in SAS, so a forward selection was conducted manually. Forward selection methods start with no variables in the model. In the first step the performances of all of the variables are evaluated individually. The variable that maximizes group separation is selected and entered. This process is repeated for the remaining variables until adding variables no longer contributes to group separation, or the partial *F*-statistic evaluated with Wilk's Λ (Rencher, 2002). The holdout classification rate was used to evaluate variable performance and selection for the KNN and QDA procedures in this analysis. Variables

were added to the analysis one-by-one, selecting the variable that most greatly improved holdout classification, until adding variables ceased to significantly improve performance. The selection was stopped after three steps yielded no significant improvement. The re-substitution, jackknife, and holdout classification rates for all variable subsets were evaluated to determine which subset to select.

Evaluation and Comparison

There are a variety of methods by which classification performance may be evaluated. Three common options are re-substitution, jackknife cross-validation, and holdout validation. The above three methods are different ways of estimating an actual error rate from the available sample and can be compared across methods to compare classification performance.

Re-substitution estimates the actual error rate by reclassifying the observations used in the estimation of the classification functions (Rencher, 2002; Manly, 2005; Huberty and Olejnik, 2006; Fielding, 2007). Though it is the simplest, it is the least accurate of the three methods, providing upwardly biased hit rate estimation. The classification rules for discriminant methods are derived to maximize between-group variation of the sample, minimizing the number of misclassifications made (Rencher, 2002; Huberty and Olejnik, 2006). Thus, estimating the classification rate using data on which the classification functions were derived provides an optimistically biased estimate.

Jackknife cross-validation, or leave-one-out cross-validation, is one method used to obtain a more accurate classification rate estimate (Manly, 2005; Huberty and Olejnik, 2006). To avoid providing an estimate based upon the classification of

observations used to derive the classification rules, it derives the classification functions using all of the observations in the sample while holding one observation out. The held out observation is then classified using the classification functions estimated from the other observations. This process is repeated for every observation in the sample and the average classification rate from all of the runs is reported. However, this can still provide an optimistically biased estimate of the classification rate as it is usually only slightly lower than that offered by re-substitution (Manly, 2005). Furthermore, jackknife validation can be unreliable because it is known to have a large variance for its classification rate estimate (Huberty and Olejnik, 2006; Fielding, 2007).

The last method of classification rate estimation, holdout validation, avoids biasing the estimate by splitting the overall sample into two smaller samples: training and holdout. The training sample is used to estimate the classification rules. The rules are then applied to the holdout sample to estimate their actual classification rate (Rencher, 2002; Huberty and Olejnik, 2006; Fielding, 2007). Though this method does well at reducing the bias in estimating the classification rate, it has a few drawbacks. First, it requires a large sample size (Rencher, 2002; Huberty and Olejnik, 2006). Second, the classification functions derived on the training sample are not the same as those that would be derived if all available data were used. Theoretically, classification rules derived from all available data would perform better than those derived from a subset of the data (Rencher, 2002; Huberty and Olejnik, 2006). Third, the size of the test set in terms of proportion of the overall sample must be determined. Huberty and Olejnik (2006) note that there has not been a lot of work to provide guidelines on how to do this, but that the training sample should be about 75% of the overall sample size.

In this comparison of discriminant method performance, out-of-sample accuracy is of particular interest. Therefore, this analysis evaluates classification performance with a holdout sample in order to obtain the least biased out-of-sample classification rate estimates possible. The training sample is 70% (n_t=855) of the overall sample, while the holdout sample is the remaining 30% (n_h=378). The observations included in each were chosen at random. Table 9 and Table 10 show the training and holdout sample sizes respectively.

Table 9. Training sample size

Group	Frequency	Percent
White males	442	51.70
White females	226	26.43
Black males	79	9.24
Black females	52	6.08
Hispanic Males	56	6.55
Total	855	100.00

Table 10. Holdout sample sizes

Group	Frequency	Percent
White males	187	49.87
White females	105	28.00
Black males	38	10.13
Black females	23	6.13
Hispanic Males	22	5.87
Total	375	100.00

CHAPTER III

RESULTS AND DISCUSSION

Results

Linear Discriminant Analysis

The first step in the linear discriminant analysis (LDA) was the variable selection. Stepwise selection was performed with a *p*-value for entry into the model set to 0.20 and a *p*-value for retention set to 0.15. The procedure selected 13 variables, all of the variables except the chords (i.e., frontal chord, occipital chord, parietal chord). The results are shown in Figure 3.

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	ZYBri		0.4043	144.20	<.0001	0.59574517	<.0001	0.10106371	<.0001
2	2	GOLri		0.1830	47.53	<.0001	0.48674746	<.0001	0.13985933	<.0001
3	3	NLBri		0.1394	34.33	<.0001	0.41891066	<.0001	0.17431096	<.0001
4	4	XCBri		0.0858	19.87	<.0001	0.38297154	<.0001	0.19274885	<.0001
5	5	NLHri		0.0494	10.99	<.0001	0.36405687	<.0001	0.20051709	<.0001
6	6	OBHri		0.0566	12.69	<.0001	0.34343388	<.0001	0.21095201	<.0001
7	7	OBBri		0.0566	12.66	<.0001	0.32399390	<.0001	0.22172646	<.0001
8	8	AUBri		0.0376	8.24	<.0001	0.31180912	<.0001	0.22913332	<.0001
9	9	MDHri		0.0366	8.00	<.0001	0.30038856	<.0001	0.23480696	<.0001
10	10	BBHri		0.0337	7.33	<.0001	0.29027117	<.0001	0.24003905	<.0001
11	11	WFBri		0.0343	7.47	<.0001	0.28030346	<.0001	0.24456508	<.0001
12	12	FOLri		0.0230	4.93	0.0006	0.27386741	<.0001	0.24759246	<.0001
13	13	BNLri		0.0115	2.43	0.0465	0.27073071	<.0001	0.24953744	<.0001
14	14	OCCri		0.0071	1.50	0.1998	0.26880216	<.0001	0.25038025	<.0001
15	13		OCCri	0.0071	1.50	0.1998				

Figure 3. SAS stepwise selection results

While the results show all 13 selected variables as significant, the partial R^2 for many of the variables is very small, indicating that they do not explain much variability. Therefore, the variables were removed one by one in backwards order from which they were selected, removing the least significant variables first, while assessing the performance of the model at each step. The classification rates were observed to see how they were affected and if a model with fewer parameters could perform as well as the 13-variable model.

The holdout performance was maintained down to a six-variable model. The holdout classification rate dropped about 6%, from 64.34% to 58.10%, upon going from the six-variable model to a five-variable model. Therefore, the former was determined to be the most parsimonious model that maintained the classification rate integrity. It was applied in testing the LDA performance along with the 13-variable model because it

had a classification rate equivalent to the model selected by the stepwise selection and was simpler. However, its cross-validation classification rate is about 5% lower. Table 11 shows these results.

Table 11. LDA classification rates in %

Subset	Re-substitution	Cross-Validation	Holdout
13 Variables	71.02	67.90	64.22
12 Variables	71.34	67.48	65.8
11 Variables	70.05	66.06	65.72
10 Variables	68.42	64.80	65.66
9 Variables	68.45	66.40	66.18
8 Variables	67.49	65.56	63.65
7 Variables	66.77	64.81	65.37
6 Variables	**64.65**	**62.87**	**64.34**
5 Variables	63.20	60.97	58.10
4 Variables	61.21	59.54	55.40

The six-variable model included bizygomatic breadth (ZYB), maximum cranial length (GOL), nasal breadth (NLB), maximum cranial breadth (XCB), nasal height (NLH), and orbital height (OBH). It correctly classified 64.34% of the holdout sample and had a jackknife classification rate of 62.87%, but the individual group classification rates differ. The model classified White females the most accurately and Hispanic males least accurately. The White male hit rate was 66.31%, the White female rate was

74.29%, the Black males was 57.89%, Black females was 62.22%, and Hispanic males was 54.55%.

C-q plots generated in SAS show the White male and White female groups do not exhibit a multivariate normal distribution on this variable subset (c-q plots for all analyses are shown in the appendix). However, the lack of normality seems as though it could be due to the presence of outliers. Nonetheless, the Box M-Test indicates that the covariance matrices are not equal across groups (X^2=171.30, p<.0001). Thus, the accurate classification of these five groups on this six-variable subset may not be best achieved with LDA. Natural log transformations were carried out on the variables to see if violations of normality and variance homogeneity could be resolved. However, neither assumption was satisfied by the transformed data.

Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) does not have stepwise variable selection procedures available. The assumptions of QDA are not the same as those of the statistics used in stepwise variable selection. Therefore, they are not appropriate for this analysis. However, variable selection was performed in a forward, step-up manner using the holdout classification rate for selecting the variable to add in each step. A three-variable model [including frontal chord (FRC), nasal breadth (NLB), and biauricular breadth (AUB)] was selected with this method, though a six-variable model (including FRC, NLB, AUB, GOL, OBH, and NLH) may have been appropriate as well. The order in which the variables are listed above is the order in which they were added to the model. The results of the forward selection are shown in Table 12.

Table 12. QDA forward selection results

Model	Re-substitution	Cross-Validation	Holdout
1 Variable	29.4	29.4	34.84
2 Variables	43.18	42.38	40.63
3 Variables	**52.64**	**49.86**	**50.80**
4 Variables	57.68	54.71	53.53
5 Variables	61.46	58.76	54.63
6 Variables	**63.98**	**58.63**	**58.96**
7 Variables	69.02	59.05	60.93
8 Variables	69.99	57.70	60.23
9 Variables	72.38	57.84	61.51

Though the six-variable model has better hit rates for all of the estimates, the three-variable model was ultimately chosen because the improvement in classification accuracy when moving to the six-variable model is not great when considering the number of additional steps it took to achieve; three more variables were added to achieve an 8% increase in classification accuracy. This can add a good deal of complexity to a quadratic equation. This consideration along with the notoriety of QDA models for over-fitting and increased sensitivity to errors drove the selection of the simpler model. Table 13 shows the classification estimates for the three-variable, six-variable, and a 16-variable model.

In this case, only the White male data are not multivariate normally distributed. Once again, the lack of normality appears to be caused by outliers. Otherwise, the

majority of the points on the c-q plot are almost perfectly in line with a multivariate normal distribution. The groups still do not display equal covariance matrices, as indicated by a Box M-Test (X^2=53.08, p<.001). Thus, as far as the assumptions go, this analysis seems appropriate for the distributional properties of the data. However, it performs poorly, only correctly classifying about 50% of the observations. The individual group classification rates are: White males – 45.99%; White females – 61.90%; Black males – 52.63%; Black females – 43.48%; and Hispanic males – 50%.

Table 13. QDA classification rates in %

Model	Re-substitution	Cross-Validation	Holdout
All Variables	80.09	53.04	51.74
3 Variable	52.64	49.86	50.80
6 Variables	63.98	58.63	58.96

K-Nearest Neighbor Analysis

K-nearest neighbor analysis (KNN) does not include stepwise variable selection procedures for the same reason as QDA: its assumptions differ from those of the variable selection statistics. Therefore, a forward selection was performed by the same method as that of the QDA. The results for the k=3 analysis are shown in Table 14.

Table 14. KNN (*k*=3) forward selection results

Model	Re-substitution	Cross-Validation	Holdout
1 Variable	36.32	28.93	34.64
2 Variables	59.22	28.27	32.78
3 Variables	76.48	39.12	46.67
4 Variables	81.72	48.12	48.39
5 Variables	**79.79**	**43.74**	**53.43**
6 Variables	80.34	43.97	52.07
7 Variables	80.64	42.45	51.88
8 Variables	**82.30**	**47.75**	**50.02**

The eight-variable model was selected as the best for this analysis. Adding more variables after five had negative effects on the holdout classification rate three steps in a row. However, the eight-variable model has a stronger cross-validation classification rate, which may indicate that it gains performance though its holdout rate is lower. Furthermore, the eight-variable model offers more stable classification rate estimates, as its cross-validation and holdout rates are not as divergent as the five-variable model.

The re-substitution rate rapidly diverged from the cross-validation and holdout rates; it immediately pulled away from the other two. The classification rates for the eight-variable model as well as an all variables model are shown in Table 15. The eight-variable model includes mastoid height (MDH), NLB, ZYB, GOL, occipital chord (OCC), BBH, FOL, and XCB. The variables were added to the model during selection in that order. The individual group holdout classification rates for the eight-variable

model are: White males – 65.24%; White females – 52.38%; Black males – 39.47%; Black females – 43.48%; and Hispanic males – 54.55%.

The results for the *k*=1 KNN are shown in Table 16. Improvement gains ceased after the addition of a second variable, so the two-variable model was chosen. Once again, the re-substitution classification rate quickly diverges from that of the cross-validation and holdout rates, even reaching 100% with only five variables in the model. The *k*=1 analysis did not perform as well as the *k*=3. It never reached the holdout classification rates of the *k*=3 models and the selected model's overall holdout classification is only 44.23%, about 10% worse than that of the *k*=3 analysis. The two variables in the model are ZYB and NLB. The model's group holdout classifications are: White males – 40.11%; White females – 66.67%; Black males – 39.47%; Black females – 52.17%; and Hispanic males – 22.73%.

Table 15. KNN (*k*=3) classification rates in %

Model	Re-substitution	Jackknife	Holdout
All Variables	89.47	46.20	45.99
8 Variable	82.30	47.75	50.02

Table 16. KNN (*k*=1) results in %

Model	Re-substitution	Jackknife	Holdout
1 Variable	45.35	30.77	34.38
2 Variable	**69.41**	**32.65**	**44.23**
3 Variable	93.37	34.67	45.25
4 Variable	99.78	38.05	45.57
5 Variable	100	41.87	46.72
6 Variable	100	39.87	48.39
All Variables	100	40.69	41.33

An assessment of multivariate normality shows the White males and Black females data are not normally distributed for the eight-variable subset used in the *k*=3 analysis. The not normal distribution of some groups may make KNN appropriate. However, the homogeneity of variance assumption is violated, indicated by a significant Box M result (X^2=99.50, *p*=0.0024). The *k*=1 analysis had similar results, but only the Black male data was normally distributed. The Box M-Test was significant, indicating variance heterogeneity (X^2=67.64, *p*<0.0001). Natural log transformations were performed in an attempt to achieve homogeneity of variance, but tests still indicated significant differences.

Classification Trees

The Classification Tree (CT) analysis was done in SAS Enterprise Miner 12.1. The Gini Index was used as the measure of impurity. The tree was restricted to binary splits, only allowing for the creation of two subsets per split. The maximum depth was

set to 10, limiting the level of node generations to 10 after the root node, which is generation 0. Lastly, the tree was grown on the training sample and assessed and selected based upon the misclassification rate of the holdout sample.

The results of the CT analyses are shown in Table 16. The final tree had 16 terminal nodes. The re-substitution and holdout (validation) rates are 74.04% and 72% respectively. However, though the overall classification rates are high, the tree's performance is not good; this tree only accurately predicts White males and White females. The other groups are not predicted well. The large discrepancy between group classification rates is due to the vastly differing sample sizes. If there are divergent sample sizes between groups included in a classification tree analysis, accurately classifying the smaller group(s) does not significantly increase the accuracy or decrease the impurity enough for the algorithm to notice. Thus, though decision trees may have good overall classification rates, they do a poor job of group classification when constructed on a training sample with differently sized groups. This is demonstrated here, where the White male group holdout classification rate is about 87% and the White female rate is about 85%. On the other hand, only 16% of Black males were correctly classified, 35% of Black females, and 18% of Hispanic males in the holdout sample. Therefore, a subsample of the data was used in another analysis in which all groups have equal sample sizes in order to obtain some indication of how well CTs could perform for ancestry estimation.

Table 17. Classification tree results

Model	Re-substitution	Validation
Full Sample	74.04	72.00
Equal Samples	60.80	52.80*

- v-fold cross-validation

A stratified random sample of the data was taken in SAS Enterprise Guide 5.1 where each group has 75 observations because that is the sample size of the smallest group in the data set: Black females. The following analysis used the same specifications as the prior CT. However, a holdout sample could not be created because the group sample sizes are too small. Therefore, v-fold cross-validation was used to assess and select the tree. This is the only method other than re-substitution offered in Enterprise Miner when a validation sample is not used. Breiman *et al.* (1984) also state that v-fold cross-validation provides accurate classification rate estimates in CTs.

The overall tree performance using the equal samples was much lower than that of the full sample; the re-substitution rate was 60.80%, whereas the full sample re-substitution rate was 74.04%. The cross-validation rate (shown in the validation column of Table 16) is also quite low. However, the group re-substitution rates for this run are more reasonable: White males – 54.67%; White females – 69.33%; Black males – 50.67%; Black females – 62.67%; and Hispanic males – 66.67%. These rates make it clear that the divergent sample sizes of the CT on the original training data affected the results.

Discussion

Though each of the applied discriminant techniques attempts to separate groups by different methods, they all found many of the same variables as significant predictors of group membership. Table 18 lists the variable selected by each method. Commonly reoccurring variables include: ZYB; XCB; GOL; and NLB. The first is mainly associated with sex discrimination, while the latter three are more likely associated with ancestry.

Table 18. Variables selected by the discriminant methods

Model	Variables
LDA	ZYB GOL NLB XCB NLH OBH OBB AUB MDH BBH WFB FOL BNL
QDA	FRC NLB AUB GOL OBH NLH
KNN	MDH NLB ZYB GOL OCC BBH FOL XCB
CT	[ZYB][1] [XCB GOL][2] [NLB BNL][3] [FOL][4]

- The variables are listed by the order in which they were selected. Bracketed variables were used for splitting different nodes that were in the same level of the tree. The superscript indicates the level of the tree, showing the order of selection.

The overall re-substitution, jackknife, and holdout classification rates for each method are displayed in Table 19, while the group specific holdout and re-substitution rates are shown in Table 20 and Table 21 respectively. The CT achieved the greatest overall holdout classification accuracy, correctly classifying 72% of the observations in the holdout sample. However, this figure is misleading, as the classification rate seems due to the large sample size difference between the groups; the CT only predicts the White male and White female groups accurately. These two groups combined account for about 78% of the training sample. This biases the classification tree because the

criteria used for evaluating splits, an impurity measure (in this case the Gini Index), is calculated based upon proportions of observations from each group that are present in the new nodes. The algorithm is always looking to make pure terminal nodes. Therefore, it is still possible to predict smaller groups. However, with groups that are difficult to cleanly separate, such as these, terminal nodes are likely to get assigned to the large groups because they will account for the greatest proportion of observations in many of the nodes. Therefore, another CT was run on a stratified random sample of the data to assess how well it might work with groups of equal size.

The equal sample CT results are shown in Table 19 and Table 21 under "CT (=n_i)". Unfortunately, the group samples sizes (n_i = 75) were too small to split into training and holdout samples, so this analysis could not be evaluated by holdout classification. Also, though SAS Enterprise Miner 12.1 would evaluate split decisions and tree performance with v-fold cross-validation, it would not return a classification matrix. Therefore, the re-substitution rates have to be used for assessing group classification rates.

Table 19. Overall classification rates in %

Model	Re-substitution	Cross-Validation	Holdout
LDA (13 var.)	71.02	67.90	64.22
LDA (6 var.)	64.65	62.87	64.34
QDA (3 var.)	52.64	49.86	50.80
QDA (6 var.)	63.98	58.63	58.96
KNN *k*=3	82.30	47.75	50.02
KNN *k=1*	69.41	32.65	44.23
CT	74.04	NA	72.00
CT (=n_i)	60.80	52.80	NA

The overall cross-validation performance for the equal-samples CT shown in Table 19 is 52.8% and the re-substitution rate is 60.8%. The classification rates for the Black males, Black females, and Hispanic males are much better, but the rates for White males and White females went down. This model is probably a more realistic representation of CT performance because differing group sample sizes did not bias the analysis. However, as previously discussed, re-substitution classification rates can offer unrealistically high estimates of model performance. Table 17 demonstrates this with the KNN rates. The LDA, QDA, and CT rates show that the re-substitution rate is not always a large overestimate. However, the 8% difference between the overall re-substitution rate and cross-validation rate for the equal samples CT indicate that the re-substitution rate may overstate the performance. Unfortunately, the degree to which it is an overestimate on individual groups cannot be evaluated without an out-of-sample classification matrix.

Table 20. Group holdout classification rates in %

Model	White Male	White Female	Black Male	Black Female	Hispanic Male
LDA (6 var.)	62.03	72.38	63.16	69.57	54.55
QDA (3 var.)	45.99	61.90	52.63	43.48	50.00
KNN *k=3*	56.15	56.19	52.63	52.17	50.00
CT	87.16	84.76	15.79	34.78	18.18

Table 21. Group re-substitution classification rates in %

Model	White Male	White Female	Black Male	Black Female	Hispanic Male
LDA (6 var.)	65.16	65.93	64.56	61.54	66.07
QDA (3 var.)	43.44	58.85	55.70	48.08	57.14
KNN *k*=3	67.19	62.39	87.34	98.09	83.83
CT	89.37	85.40	22.78	26.92	23.21
CT (=n_i)	54.67	69.33	50.67	62.67	66.67

While CT did not perform as well as LDA its performance was not far off, based upon the available classification estimates in Table 16. Furthermore, while it does have reduced performance, it has many advantages that simplify analysis and interpretation as well as reducing uncertainty of the results. CT makes no assumptions pertaining to data distributions. Thus, assumptions need not be tested and there are no violations that can invalidate results, compromise performance, or complicate interpretation. Furthermore, these conditions remain consistent regardless of which and how many

groups and variables are included. CT, by nature of its workings, has built-in variable selection and is robust to outliers, errors, and misclassifications in the data. Furthermore, they have a variety of methods available for handling missing data and do not require the deletion of observations with missing values (Breiman *et al.*, 1984; Feldesman, 2002; Ding and Simonoff, 2010). This would allow retention of more observations, which is particularly important when one of the smaller groups (e.g., Hispanic males, Hispanic females, Black females) are involved. Furthermore, it could make some of the other variables with larger degrees of missingness more useful, alleviating the degree to which sample sizes would be reduced due to their inclusion in an analysis. Eight variables were removed from the data set in this analysis and FORDISC flags variables with large proportions of missing values because of the limitations their inclusion places on sample sizes. Therefore, a method that does not require the deletion of observations with missing values may be very beneficial in this context.

LDA and QDA both had consistent estimates of classification performance; their re-substitution, cross-validation, and holdout estimates are close to one another, with a maximum of a 2-3% difference between them. However, LDA outperformed QDA by about 14% in the overall holdout rate. LDA's individual group classification rates were also greater for every group. Therefore, LDA should be the preferred method for classification when deciding between these two.

Posterior probabilities of classification for correctly classified and misclassified observations provide additional evidence of LDA's improved performance over QDA (Figures 30 to 37 in the appendix show histograms and descriptive statistics). LDA's

correctly classified observations tend to have greater posterior probabilities of belonging to the group into which they were classified. QDA's misclassified observations tend to have lower posterior probabilities of classification, but the majority of misclassified observations for LDA and QDA are not near misses; LDA's median posterior probability for its misclassifications is in the mid 0.5s, while QDAs is in the mid 0.4s. The misclassification posterior probability distributions for each method are skewed to the right; most observations lie on the lower end with relatively few larger values.

Observing the difference between the maximum posterior probability of classification and the posterior probability of classification into the correct group on misclassified observations (Figures 38 to 55 in the appendix) provides an image of how many observations were near misses. Observing quartile statistics (Figures 39, 42, 45, 48, 51, and 54) reveals that only about 25% of them had posterior probabilities of belonging to the proper group that were about 0.14 less than that of the maximum posterior probability. While the statistics for the first quartile are very similar for LDA and QDA, the median difference tended to be in the mid-0.30s for LDA and mid-0.20s for QDA. This confirms that QDAs misses were nearer than LDAs, but neither method's misclassified observations are mostly near misses.

Researchers should be cautious when using LDA due to the violations of its assumptions, which may complicate interpretations of significance (Manly, 2005) and call into question whether the models can be generalized to out of sample observations (Ousley and Jantz, 2012). It should also be recognized that LDA does not perform optimally under conditions violating its assumptions (Huberty and Olejnik, 2006). While QDA's assumptions were better upheld, caution is also warranted when applying it.

Holden and Kelley (2010) note QDA's increased sensitivity to misclassifications and errors in the training data. Furthermore, QDA is known for developing over-fit models that do not perform as well on out of sample data. Thus, models developed on available data may be optimistic when applied to new instances.

A significant advantage of LDA, is the availability of variable selection techniques that may find a subset of the original variables that most accurately predicts group. This simplifies the analysis and interpretation by reducing dimensionality. It also leads to the creation of more parsimonious models, which are generally more accurate in a predictive analysis (Huberty and Olejnik, 2006). However, it is not always the case that the most parsimonious model is found by variable selection or that the selected variables are actually significant predictors of group membership. Manly (2005) explains that, given enough variables, it is possible for variable selection procedures to find a significant subset simply by chance. This analysis was a good demonstration of how a greater number of variables can be selected than are necessary. The stepwise selection found a subset of 13 significant variables. However, the selection statistics showed that the partial R^2 for many of the variables was very small. Removing the variables one by one in a backward manner while observing the classification rates resulted in the selection of a six-variable model that performed equally as well as the 13-variable model on the holdout. However, observation of the cross-validation may lead to the selection of a 12 or 11 variable model. While the 6-variable model is more parsimonious, the 13-variable model classification rates do not suggest that it is over-fit; the various classification rate estimates are reasonably close. It is also possible for

variables that are not significant in the selection to contribute to classification. Careful attention should be paid to the results of a variable selection for these reasons.

KNN analysis had erratic classification estimates. The re-substitution rate rapidly rose with the addition of new variables, while the cross-validation and holdout rates were not as greatly affected. The cross-validation rate is also lower than the holdout rate in both the k=3 and k=1 cases. The strange patterns and behavior of the classification rates are cause for concern. Typically, large differences between internal and out-of-sample classification rates are indicative of an over-fit model. In this case it seems that KNN models can become over-fit very rapidly. Large differences between the re-substitution and jackknife and holdout rates occur with a two-variable model. Furthermore, holdout rates do not rise very high before beginning to fall with either KNN model.

It is not clear as to why there is such a great difference between the re-substitution and leave-one-out cross-validation rates. KNN does not construct functions for classification, it classifies by observing group membership of nearby observations. Thus, if leave-one-out cross-validation is as simple as taking an observation out of the data set and then reclassifying it based upon the training sample, then the re-substitution and cross-validation classification rates would be identical, or at least nearly so; upon being placed back into the training sample, the left out observation should have the same nearest neighbors as it did when creating the re-substitution classifications. This is especially true in SAS because it calculates distances using a covariance matrix that includes the left out observation. Unfortunately, the SAS

documentation is not clear on exactly how it performs leave-one-out cross-validation for the k-nearest neighbor procedure.

Dimensionality is one possible explanation for the large difference between the re-substitution and cross-validation classification rates. Beyer *et al.* (1999) revealed that as dimensionality increases, the distance from an observation to its nearest neighbor approaches the distance between that observation and its furthest neighbor. It was found that these distances decrease the fastest within the first 20 dimensions and that having as few as 10 dimensions can cause the nearest neighbor procedure to become unstable, making the distance between the nearest neighbor and other data points negligible. While only eight variables were used here, this number still causes significant reduction in the difference between nearest and furthest neighbor distance. Furthermore, Beyer *et al.* (1999) reported that smaller values of *k* are more sensitive to dimensionality, and a small value of *k* (3) was used. However, Table 14 shows that the large difference between the re-substitution and cross-validation rates occurs with only three variables, so dimensionality is not necessarily the issue. The gap between the two estimates is maintained for the model with all 16 variables; it does not become much worse after the addition of the third variable.

While KNN is more appropriate in situations where normality assumptions are violated, it may not be a very useful method in the context of forensic ancestry estimation. First, it is a nonparametric statistic that is good for situations where data are not normally distributed. In this analysis the data for many of the variable subsets ended up approximating normality on most of the groups. In some cases where normality was violated it appeared that a few outliers were the cause. Second, early

explorations of human craniometric variation noted large degrees of overlap between geographic groups (Howells, 1989). This observation has been repeated in more recent analyses (Roseman and Weaver, 2004). Therefore, it may not be the case that an individual's nearest neighbor(s) in multivariate space is from their same class. A larger *k* may improve classification if it would generally allow more observations from the same class to be considered. However, a larger *k* enables the consideration of less similar observations when classifying an unknown (Rencher, 2002; Huberty and Olejnik, 2006).

The Hispanic group had particularly low classifications in the KNN cross-validation. Table 30 and Table 32 (in the appendix) show cross-validation classification rates for the 5-variable and 8-variable KNN analyses respectively. The former only classifies about 21% of Hispanics correctly, allocating 41% of them to the White group. Table 32 shows similar figures, with 27% of Hispanics correctly classified and 45% classified into the White group. These classification rates seem to be particularly low. However, Spradley *et al.* (2008) report only 45% of Hispanics were correctly classified when four racial groups were considered in an LDA. While this exceeds the cross-validation estimates, it is closer to and lesser than the KNN holdout estimates shown in tables 31 and 33 in the appendix.

Another issue with KNN is it assumes homogeneity of variance. Every evaluation of covariance matrix equality resulted in significance regardless of how many or few variables were included. It is difficult for this assumption to be met, but it may be easier to satisfy in an analysis with fewer groups. Furthermore, it is difficult to evaluate KNN assumptions because of the distributional situation for which it is most appropriate – equal covariance matrices and lack of multivariate normality; tests for equality of

covariance matrices are known for their sensitivity to violations of their multivariate normal assumptions (Rencher, 2002). Tests for homogeneity of variance that are robust to departures from normality have been proposed (O'Brien, 1992), but they are neither common nor built into any available software known by the author.

Aside from comparisons between methods, it should be noted that none of the analyses had strong classification rates. The greatest holdout classification rate was about 64% with the next greatest being about 53%. While all classification rates were greater than random (20% in this case), none were particularly strong. Spradley *et al.* (2008) demonstrate that including fewer groups can help improve classification rates and Ousley and Jantz (2005) provide guidelines for narrowing down group membership. They state, "classifications into two to five groups are expected to be more accurate than those involving many more groups…" (Ousley and Jantz, 2005:np), but this analysis shows that five groups can still be too many for reliable classifications.

One cause of the lower classification rates could be that this analysis is estimating both ancestry and sex simultaneously; an observation is considered misclassified if it is classified into the correct ancestry group, but incorrect sex group. A crude way of observing where misclassifications more commonly lie is to aggregate ancestry groups and sex groups and observe the new classification rates. Tables with these classification rates are shown in the appendix (Table 22 to Table 45). From these tables it can be seen that ancestry is more difficult to capture than sex; overall classification rates for ancestry groups tend to be about 10% to 20% less than classification rates for sex groups. However, by pooling ancestry groups overall classification rates rise by about 8% to 15% depending on the discriminant method.

CONCLUSION AND RECOMMENDATIONS

This research was a comparative evaluation of four statistical discriminant analysis techniques in forensic ancestry estimaton: linear discriminant analysis (LDA); quadratic discriminant analysis (QDA); k-nearest neighbor analysis (KNN); and classification trees (CT). Performance evaluations were based upon correct classification rates for each method. Data consisted of 16 craniometric variables and 1257 individuals from the Forensic Data Bank were used. The data included American White, American Black, and Hispanic males and females. However, the Hispanic female group was excluded due to its small sample size.

Overall, LDA performed the best, indicated by the classification rate estimates in Table 16 and Table 19. While the classification rates for the CT in Table 16 are better than those of LDA, the results were likely due to the great disparity between group sample sizes. A follow-up CT with equal group sample sizes did not perform as well based upon the re-substitution and cross-validation estimates. QDA and KNN analysis did not perform very well, only correctly classifying 50-53% of the holdout observations. Furthermore, they have some risky drawbacks, such as QDA's notoriety for over-fitting and sensitivity to error and KNN's rapid over-fitting in this analysis. Furthermore, neither QDA nor KNN have any variable selection methods available.

LDA, however, while performing the best, also has some drawbacks. Its distributional assumptions of multivariate normality and equality of covariance are difficult to satisfy. This analysis shows that multivariate normality may be achieved by removing some outliers. However, there must be theoretical reasons behind removing

outliers; making the data fit the distributional assumptions of a particular method is not a valid reason. For example, an outlier was removed from this analysis due to a clear data entry or recording error (MDH = 113 mm.). On the other hand, homogeneity of variance does not seem as easily met, as every test in this analysis resulted in significance with very small *p*-values. Furthermore, whether or not the data meet these assumptions varies from analysis to analysis because the number of variables and groups as well as which are included will have effects on data distributions. Also, one of LDA's significant advantages – the availability of variable selection procedures – shares these assumptions and neither the LDA nor the variable selection may perform optimally with their assumptions violated.

In their discussion of LDA's assumptions, Ousley and Jantz (2012:317) state, "if groups show very different levels of variability…other statistical procedures…may be necessary, although they may not produce the most accurate statistics." The results of this analysis support their statement; the nonparametric procedures evaluated here did not perform as well as LDA. However, it should be recognized that LDA's distributional assumptions, their probable violation, difficult evaluation, and variation in whether they are met affected by variable and group selection introduce uncertainty that cannot be accounted for. This is especially true without a holdout sample with which to evaluate out-of-sample performance.

If the conditions are such that a nonparametric method should be applied, then CT is recommended. QDA is advantageous in that it does not assume variance homogeneity. However, it is sensitive to errors in the training data and is known for developing over fit models that do not perform well when applied to data not used

during model estimation. KNN is not recommended because it assumes variance homogeneity and the KNN models in this analysis became over-fit very quickly. CT, on the other hand, has many advantages among which are easy comprehension, interpretation, and application. However, users must be careful when applying them to groups with very different sample sizes.

The CT accuracy in this analysis was not far below that of LDA. Further work with CTs is recommended. This analysis only tested one set of parameter specifications that is most in line with the CART algorithm developed by Breiman *et al.* (1984). However, other settings or algorithms may provide improved performance. The vast array of benefits offered by CTs make them a strong alternative even in the face of their lower accuracy. CT does not make any assumptions about the data that if violated can confound interpretation or call into question the validity of any statistics or estimates made. Thus, there is no need for additional statistical tests of assumptions either. There is also no need for a variable selection analysis because variable selection is inherent to CT algorithms. CTs can incorporate continuous, categorical, or both variable types in its analysis. They can include observations with missing values and have a variety of methods for handling missingness. Lastly, CTs are more easily applied and interpreted than other discriminant analysis methods.

Missing values were significant issue in this analysis. LDA, QDA, and KNN cannot handle missing values. Case deletion is often used to address this, deleting observations with a value missing for any of the variables included in the analysis. It is a quick and very simple method. However, it can limit sample sizes and bias results. Furthermore, case deletion has been recognized as a poor method for handling missing

values in discriminant analyses (Acuña and Rodriguez, 2004). Case deletion is the method used in FORDISC. Even if this method does not typically produce biased results in ancestry estimations, it can significantly limit sample size depending on which variables are included. As could be seen from this analysis, some group sample sizes in the FDB are very small to begin with. These groups, if included in an analysis, cannot afford to lose many observations. Furthermore, eight variables had to be removed from this analysis because they had degrees of missingness that were too large to handle with simple missing value methods or too large to address with missing value imputation at all.

If researchers would like to continue using LDA it is recommended that some form of missing value imputation is used. FORDISC would benefit from including an imputation method rather than using case deletion. Iterative regression worked well in this analysis and did not seem heavily affected by outliers or violations of its assumptions. However, if these issues are of concern robust regression may be used. Further research could also be done to evaluate more complex methods, such as multiple imputation, that may allow for imputation of values on some of the variables with very large degrees of missingness.

Lastly, forensic ancestry estimations may benefit from the availability of more variables. FORDISC recently added three more variables to its original set of 24 (i.e., biasteronic breadth, mid-orbital width, and zygomaxillary breadth) (Ousley and Jantz, 2005). However, there are still more available that may be useful for estimating ancestry. In addition to FORDISC craniometric variables, Spradley *et al.* (2008) used variables other than those included in FORDISC and their stepwise selection found

some of these to be significant. Ancestry estimation may benefit from the evaluation and inclusion of additional variables. If other craniometric variables do not provide improved performance, they would at least provide more variables that could be useful for imputing missing values in those that are significant for ancestry analyses or for use in any of the CT methods for handling missingness.

Ultimately, LDA yielded the highest classification rates of all the tested methods. It does not seem to be greatly affected by the divergent group sample sizes and variable selection techniques may be used with it. However, its drawbacks should be recognized. The results of this analysis indicate that its assumptions are difficult to satisfy with craniometric data, especially homogeneity of variance. If tests for equality of variance are significant and a nonparametric statistic is required, then CT is recommended. Its performance was similar to that of QDA. CTs however can be affected by greatly differing sample sizes, which can be present in ancestry estimation using the Forensic Data Bank data. On the other hand, CT comes with many benefits, among which is the complete lack of distributional assumptions. KNN is not recommended due to its erratic and greatly differing classification rates. Furthermore, its assumptions are violated by these data and KNN models rapidly became over-fit in this analysis. While QDA's assumptions were satisfied by the data and its classification rates were stable it is also not recommended for use. QDA has been noted for its sensitivity to errors as well as its tendency to create models that do not perform well on data not used in their estimation.

Therefore, due to its performance, LDA should continue to be used except when violations of its assumptions suggest otherwise. In that case CT should be used.

Further work should be conducted to see if CT performance could be improved. If CT performance at least matches that of LDA, then it should certainly be used as the primary method for ancestry estimation due to its many significant advantages.

LIST OF REFERENCES

Daubert v. Merrell dow pharmaceuticals, 509 U.S. 579 (1993).
Kumho tire company, ltd. V. Carmichael, 509 U.S. 137 (1999).
Acuña E, and Rodriguez C. 2004. The treatment of missing values and its effect on classifier accuracy. Classification, Clustering, and Data Mining Applications:639-647.
Algee-Hewitt BFB. 2011. If and how many 'races'? The application of mixture modeling to world-wide human craniometric variation. [dissertation]. Knoxville, TN: University of Tennesse. 814 p.
Ayers HG, Jantz RL, and Moore-Jansen PH. 1990. Giles and elliot race discriminant functions revistited: A test using recent forensic cases. In: Gill GW, and Rhine S, editors. Skeletal attribution of race. Albuquerque, NM: Maxwell Museum of Anthropology.
Beyer K, Goldstein J, Ramakrishnan R, and Shaft U. 1999. When is “nearest neighbor” meaningful? 1540:217-235.
Birkby WH. 1966. An evaluation of race and sex identification from cranial measurments. American Journal of Physical Anthropology 24(21-28):21-28.
Brace CL. 1995. Region does not mean "race"—reality versus convention in forensic anthropology. Journal of Forensic Sciences 40(2):171-175.
Breiman L, Friedman JH, Olshen RA, and Stone CL. 1984. Classification and regression trees. Belmont, CA: Wadsworth.
Christensen AM. 2004. The impact of daubert: Implications for testimony and research in forensic anthropology (and the use of frontal sinuses in person identification). Journal of Forensic Sciences 49(3):1-4.
Connally M. 2004. Identifying covariance deifferences in comparisons of linear versus quadratic classification rule [dissertation]. Athens: University of Georgia.
Corp. I. 2011. Ibm spss statistics 20. 20.0.0.1 ed.
Ding Y, and Simonoff JS. 2010. An investigation of missing data methods for classification trees applied to binary response data. 11:131-170.
Elliott M, and Collard M. 2009. Fordisc and the determination of ancestry from cranial measurements. 5(6):849-852.
Feldesman M. 2002. Classification trees as an alternative to linear discriminant analysis. American Journal of Physical Anthropology 119:257-275.
Fielding AH. 2007. Cluster and classification techniques for the biosciences. New York: Cambridge University Press.
Fisher RA. 1936. The use of multiple measurements in taxonomic problems. 7(2):179-188.
Fisher TD, and Gill GW. 1990. Application of the giles-elliot discriminant function formulae to a cranial sample of northwestern plains indians. In: Gill GW, and Rhine S, editors. Skeletal attribution of race. Albuquerque, NM: Maxwell Museum of Anthropology.
Giles E, and Elliot O. 1962. Race identification from cranial measurements. Journal of Forensic Sciences 7(2):147-157.
Goodman AH. 1997. Bred in the bone? The Sciences 21:20-25.
Grivas CR, and Komar DA. 2008. Kumho, daubert, and the nature of scientific inquiry: Implications for forensic anthropology. Journal of Forensic Sciences 53(4):771-776.
Hintze J. 2007. Ncss. Kaysville, UT: NCSS, LLC.
Holden JE, and Kelley K. 2010. The effects of initially misclassified data on the effectiveness of discriminant function analysis and finite mixture modeling. Educational and Psychological Measurement 70(1):36-55.

Howells WW. 1973. Cranial variation in man. Papers of the Peabody Museum of Archaeology and Ethnology. Cambridge, MA. p 1-259.
Howells WW. 1989. Skull shpaes and the map. Papers of the Peabody Museum of Archaeology and Ethnology. Cambridge, MA. p 1-189.
Hron K, Templ M, and Filzmoser P. 2010. Imputation of missing values for compositional data using classical and robust methods. Computational Statistics & Data Analysis 54(12):3095-3107.
Huberty CL, and Olejnik S. 2006. Applied manova and discriminant analysis. Hoboken, NJ: John Wiley & Sons.
Iscan MY. 1988. Rise of forensic anthropology. Yearbook of Physical Anthropology 31:203-230.
Jantz RL. 2001. Cranial change in americans: 1850-1975. Journal of Forensic Sciences 46(4):784-787.
Jantz RL, and Jantz LM. 2000. Secular change in craniofacial morphology. American Journal of Human Biology 12:327-338.
Jantz RL, and Moore-Jansen PH. 1987. A data base for forensic anthropology. Knoxville, TN: The University of Tennessee.
Jantz RL, and Ousley SD. 1993. Fordisc 1.0: Computerised forensic discriminant functions. Knoxville, TN: University of Tennessee.
Khattree R, and Naik DK. 2000. Multivariate data reduction and discrimination with sas software. Cary, NC: SAS Institute Inc.
Konigsberg LW, Algee-Hewitt BFB, and Steadman DW. 2009. Estimation and evidence in forensic anthropology: Sex and race. American Journal of Physical Anthropology 139(1):77-90.
Lieberman L, Kirk RC, and Littlefield A. 2003. Perishing paradigm: Race - 1931-99. American Anthropologist 105(1):110-113.
Little RJA, and Ruban DB. 2002. Statistical analysis with missing data. Hoboken, NJ: John Wiley & Sons Inc.
Mahalanobis PC. 1936. On the generalized distance in statistics. Proceedings of the National Academy of Sciences of India 2(1):49-55.
Manly BFJ. 2005. Multivariate statistical methods: A primer. New York, NY: Chapman & Hall/CRC.
McCarty TJ. 2005. Estimation of missing data. Knoxville: University of Tennessee.
Moore-Jansen PH, Ousley SD, and Jantz RL. 1994. Data collection procedures for forensic skeletal material. Knoxville, TN: The University of Tennessee Forensic Anthropology Series.
O'Brien PC. 1992. Robust procedures for testing equality of covariance matrices. Biometrics 48:819-827.
Ousley S, Jantz R, and Freid D. 2009. Understanding race and human variation: Why forensic anthropologists are good at identifying race. American Journal of Physical Anthropology 139:68-76.
Ousley SD, and Jantz RL. 2005. Fordisc 3.0: Personal computer forensic discriminant functions. Knoxville: University of Tennessee, Knoxville.
Ousley SD, and Jantz RL. 2012. Fordisc 3 and statistical methods for estimating sex and ancestry. In: Dirkmaat DC, editor. A companion to forensic anthropology. Malden, MA: Wiley-Blackwell. p 311-329.

Pearson K. 1926. On the coefficient of racial likeness. 18(1):105-117.
Relethford JH. 1994. Craniometric variation among modern human populations. American Journal of Physical Anthropology 95:53-62.
Relethford JH. 2002. Apportionment of global human genetic diversity based on craniometrics and skin color. American Journal of Physical Anthropology 118(4):393-398.
Rencher AC. 2002. Methods of multivariate analysis. Hoboken, NJ: John Wiley & Sons.
Roseman CC, and Weaver TD. 2004. Multivariate apportionment of global human craniometric diversity. American Journal of Physical Anthropology 125:257-263.
Sauer NJ. 1992. Forensic anthropology and the concept of race: If races don't exist, why are forensic anthropologists so good at identifying them? Social Science and Medicine 34(2):107-111.
Sauer NJ, and Wankmiller JC. 2009. The assessment of ancestry and the concept of race. In: Blau S, and Ubelaker DH, editors. Handbook of forensic anthropology and archaeology. Walnut Creek, CA: Left Coast Press. p 187-200.
Smay D, and Armelagos G. 2000. Galileo wept: A critical assessment of the use of race in forensic anthropology. Transforming Anthropology 9(2):19-29.
Smith HF. 2009. Which cranial regions reflect molecular distances reliably in humans? Evidence from three□dimensional morphology. American Journal of Human Biology 21(1):36-47.
Smith HF, Terhune CE, and Lockwood CA. 2007. Genetic, geographic, and environmental correlates of human temporal bone variation. American Journal of Physical Anthropology 134(3):312-322.
Snow CC, Hartman S, Giles E, and Young FA. 1979. Sex and race determination of crania by calipers and computer: A test of the giles and elliot discriminant functions in 52 forensic cases. Journal of Forensic Sciences 24(2):448-460.
Spradley MK, Jantz RL, Robinson A, and Peccerelli F. 2008. Demographic change and forensic identification: Problems in metric identification of hispanic skeletons. Journal of Forensic Sciences 53(1):21-28.
Steadman DW, Adams BJ, and Konigsberg LW. 2006. Statistical basis for positive identification in forensic anthropology. American Journal of Physical Anthropology 131(1):15-26.
Stewart TD. 1979. Essentials of forensic anthropology. Springfield, IL: Charles C. Thomas.
SWGANTH (last updated Feb. 1, 2012). http://www.swganth.org/ (accessed Feb. 6, 2012).
Team RC. 2012. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Templ M, Alfons A, Kowarik A, and Prantner B. 2012. Vim: Visualization and imputation of missing values. R package version 3.0.1 ed.
Templ M, Kowarik A, and Filzmoser P. 2011. Iterative stepwise regression imputation using standard and robust methods. Computational Statistics and Data Analysis 55(10):2793-2806.
Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, and Altman RB. 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17(6):520-525.
Twedt DJ, and Gill DS. 1992. Comparison of algorithms for replacing missing data in discriminant analysis. Communications in Statistics: Theory and Methods 21(6):1567-1578.

Williams FLE, Belcher RL, and Armelagos GJ. 2005. Forensic misclassification of ancient nubian crania: Implications for assumptions about human variation. Current Anthropology 46(2):340-346.
Wright R. 1992. Correlation between cranial form and geography in homo sapiens: Cranid-a computer program for forensic and other applications. Archaeology in Oceania:128-134.
Wright R. 2008. Detection of likely ancestry using cranid. In: Oxenham M, editor. Forensic approaches to death, disaster, and abuse. Bowden Hills, Australia: Australian Academic Press. p 111-122.

APPENDIX

The appendix contains tables and figures referenced in the text. Figure 4 to Figure 29 are c-q plots. These provide visual evaluations of multivariate normality. C-q plots were only generated for the tests that assume multivariate normality. Table 22 to Table 45 are re-substitution, cross-validation, and holdout classification estimates aggregated for ancestry and sex for each method. These help elucidate which factor, ancestry or sex, was associated with the most classification error. Figures 30 to 37 show posterior probability descriptive statistics and histograms for LDA and QDA. The histograms are shown for correctly classified and misclassified observations. This provides an image of the strength of classification for each observation. Lastly, Figure 38 to Figure 55 show descriptive statistics, quantiles, and histograms for the difference between the maximum posterior probability (associated with the group to which an observation is classified) and the posterior probability associated with the group to which an observation belongs for LDA and QDA. This is done only for misclassified observations to evaluate how great the misclassifications were.

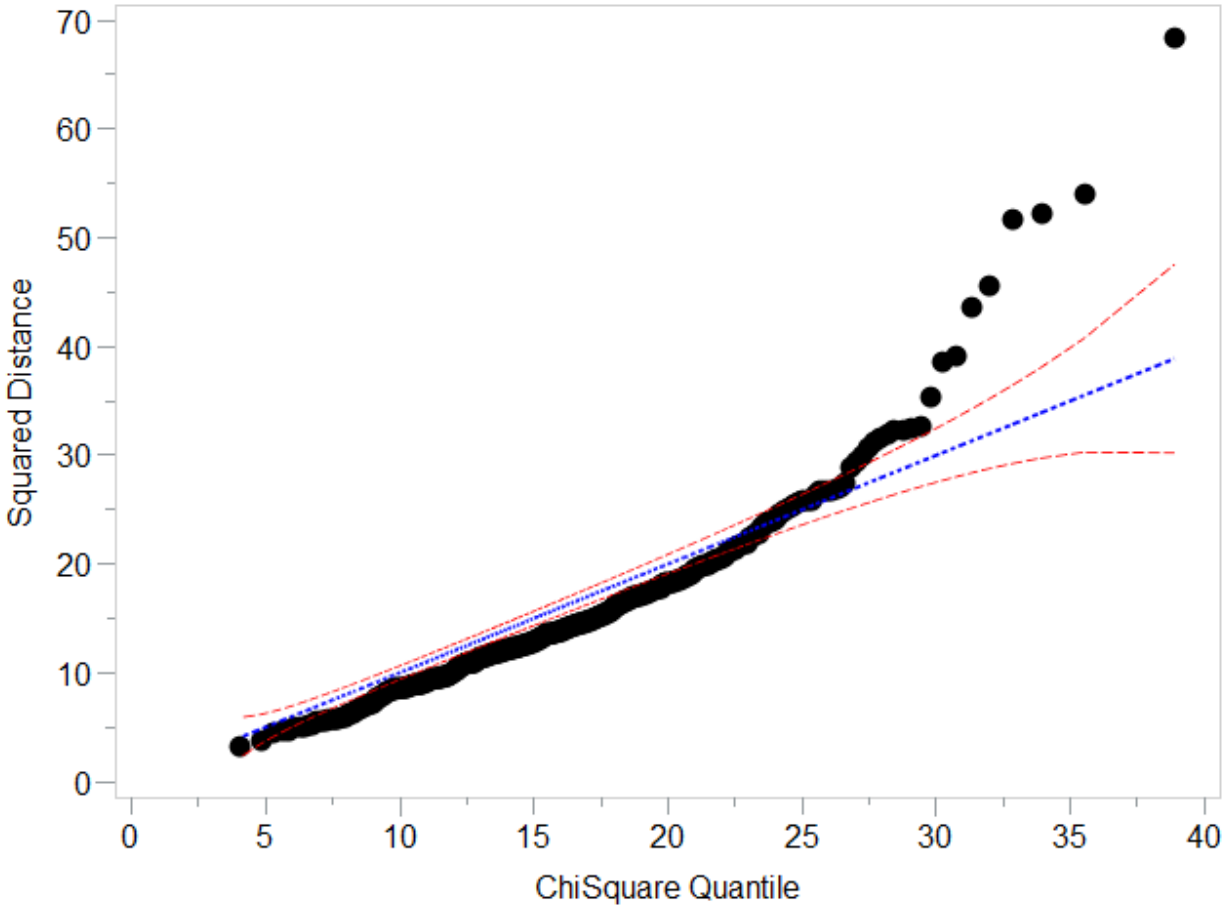

Figure 4. All variables White male c-q plot

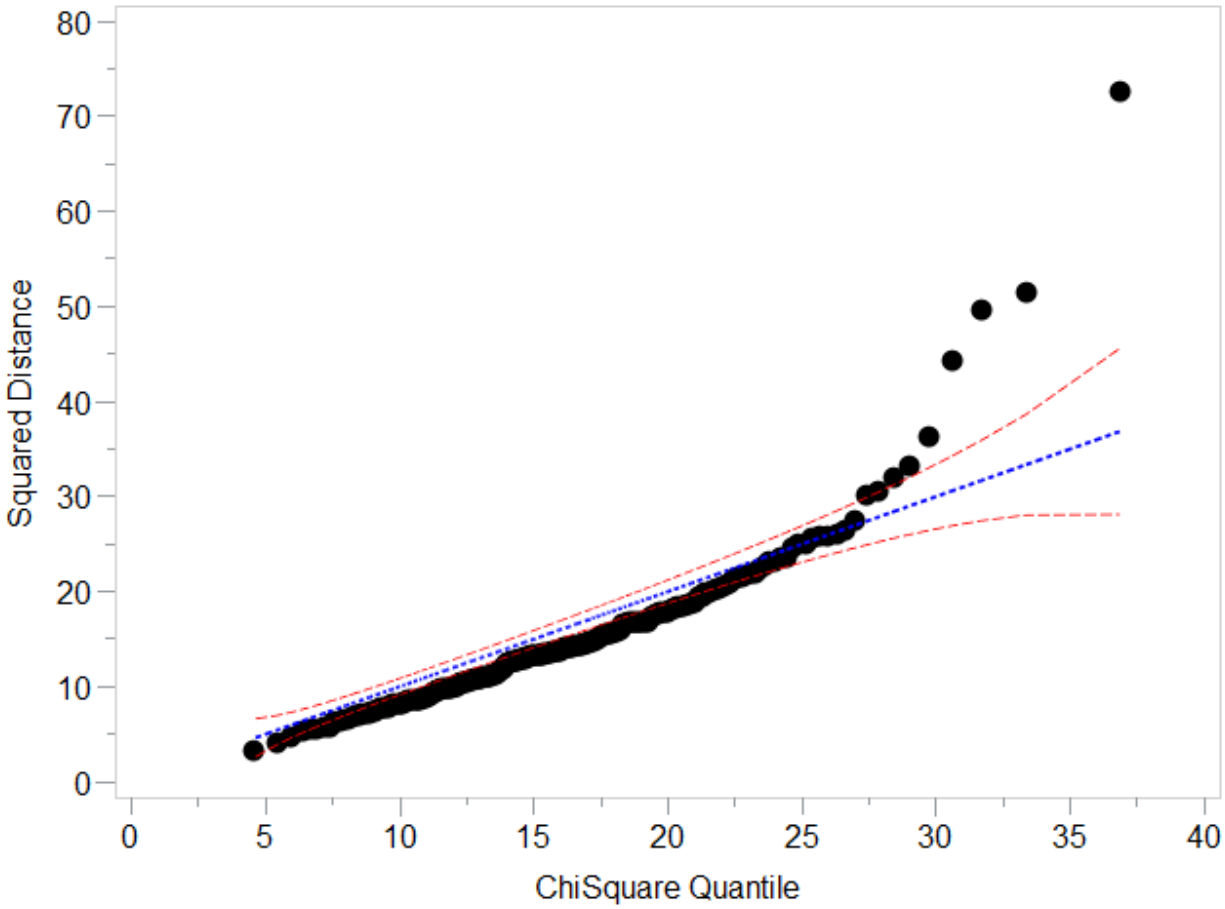

Figure 5. All variables White female c-q plot

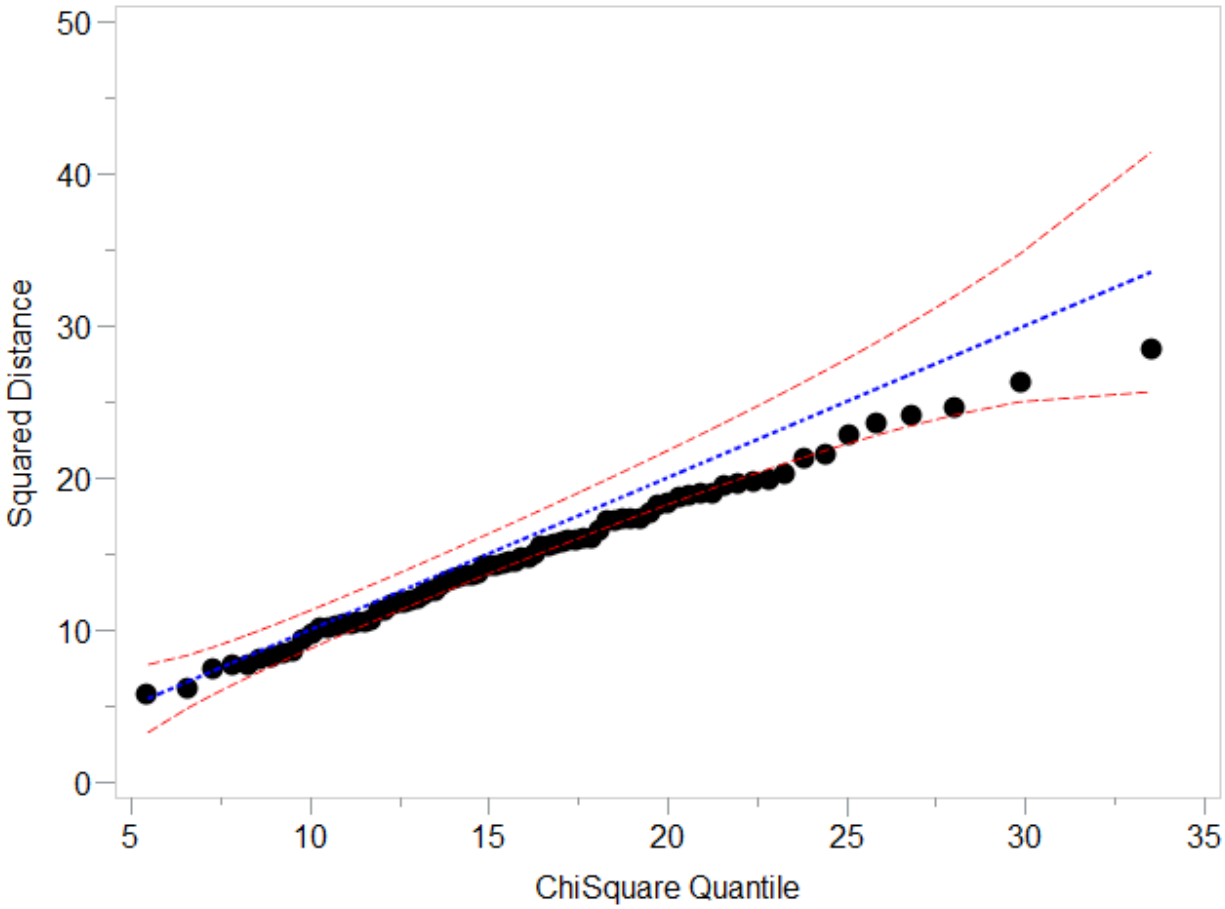

Figure 6. All variables Black male c-q plot

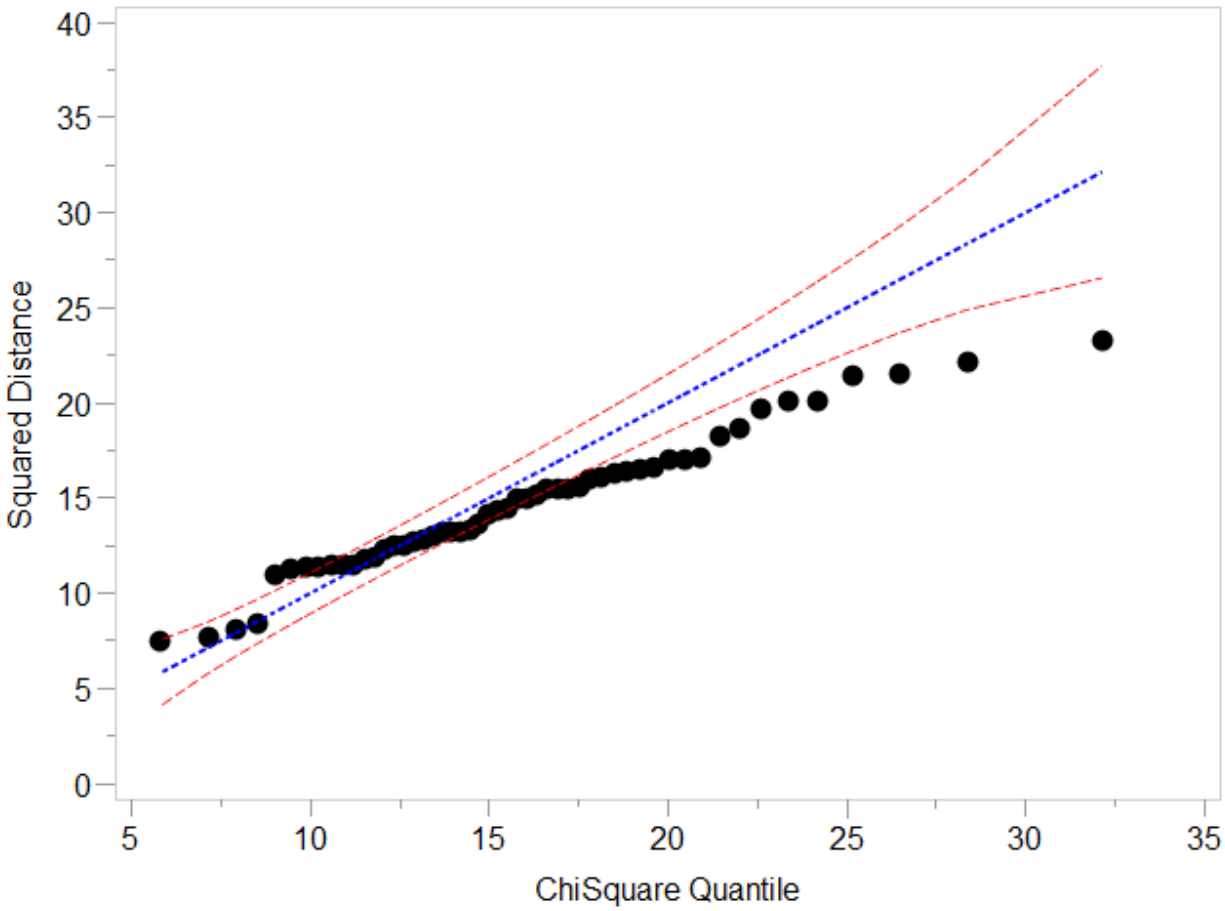

Figure 7. All variables Black female c-q plot

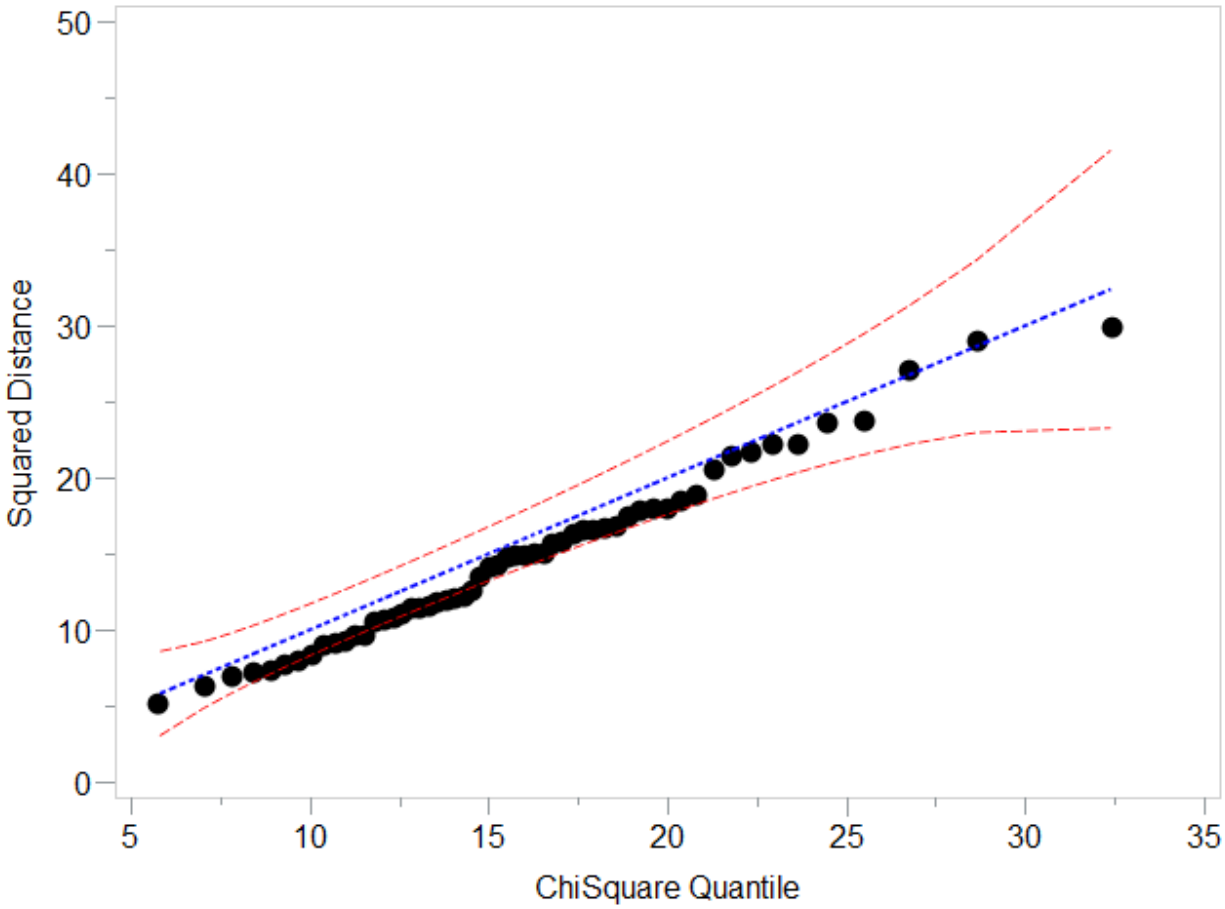

Figure 8. All variables Hispanic male c-q plot

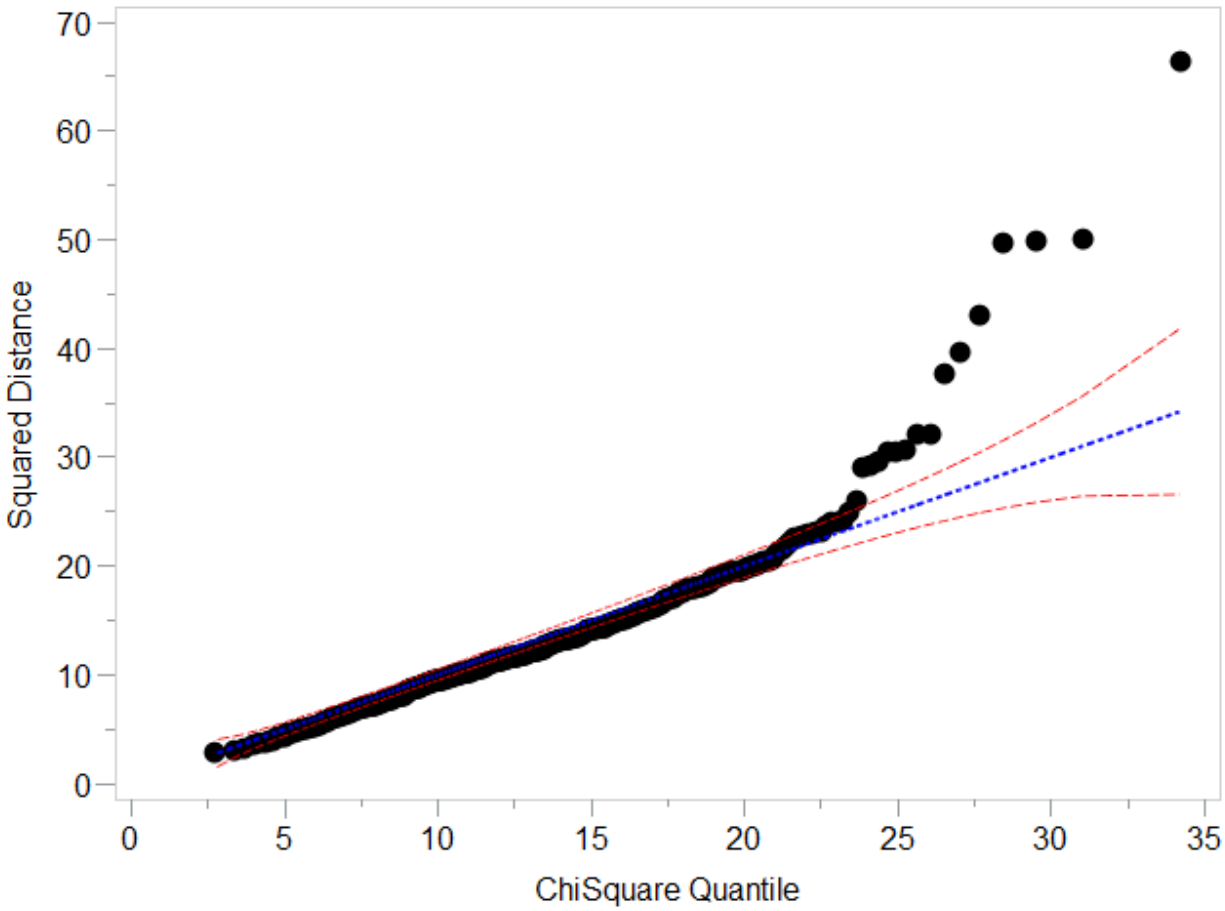

Figure 9. 13-variable LDA White male c-q plot

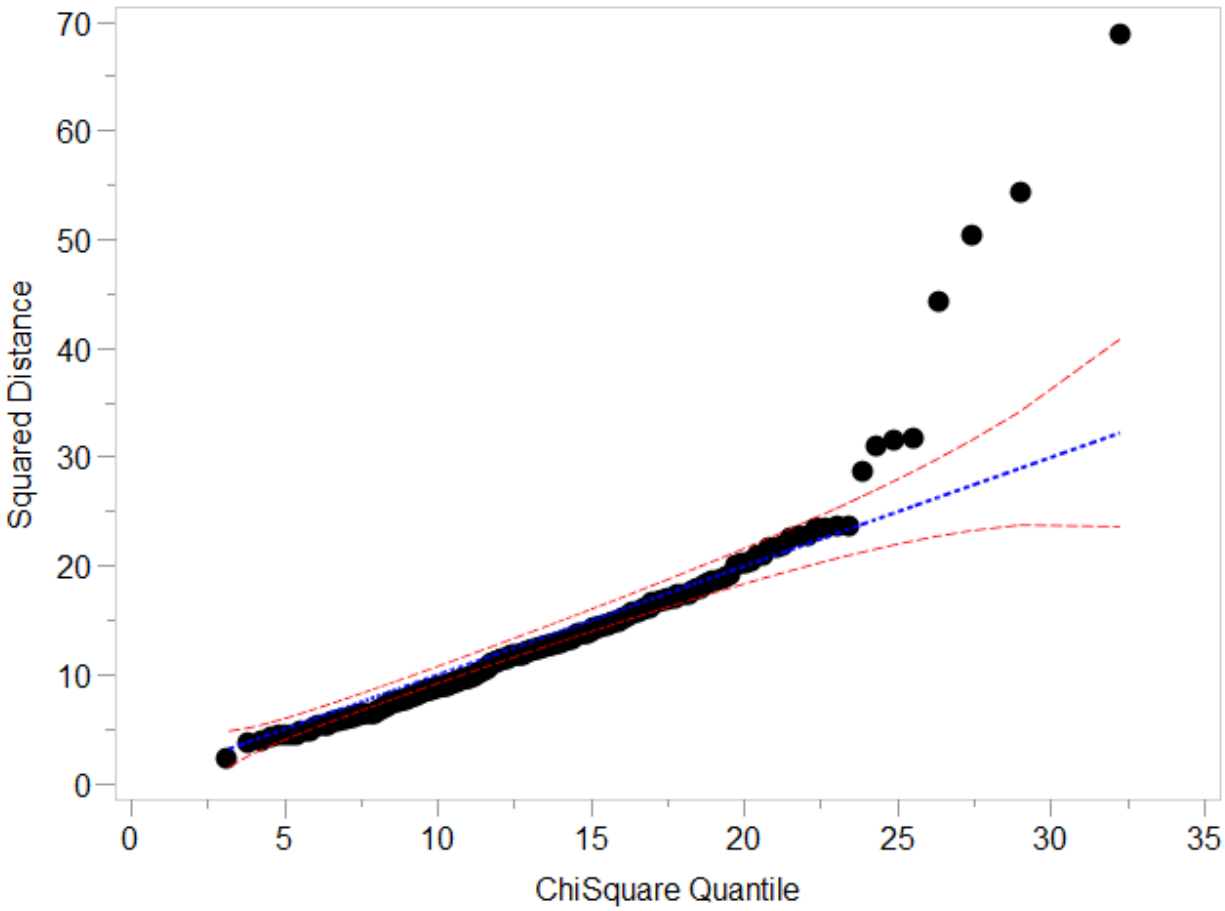

Figure 10. 13-variable LDA White female c-q plot

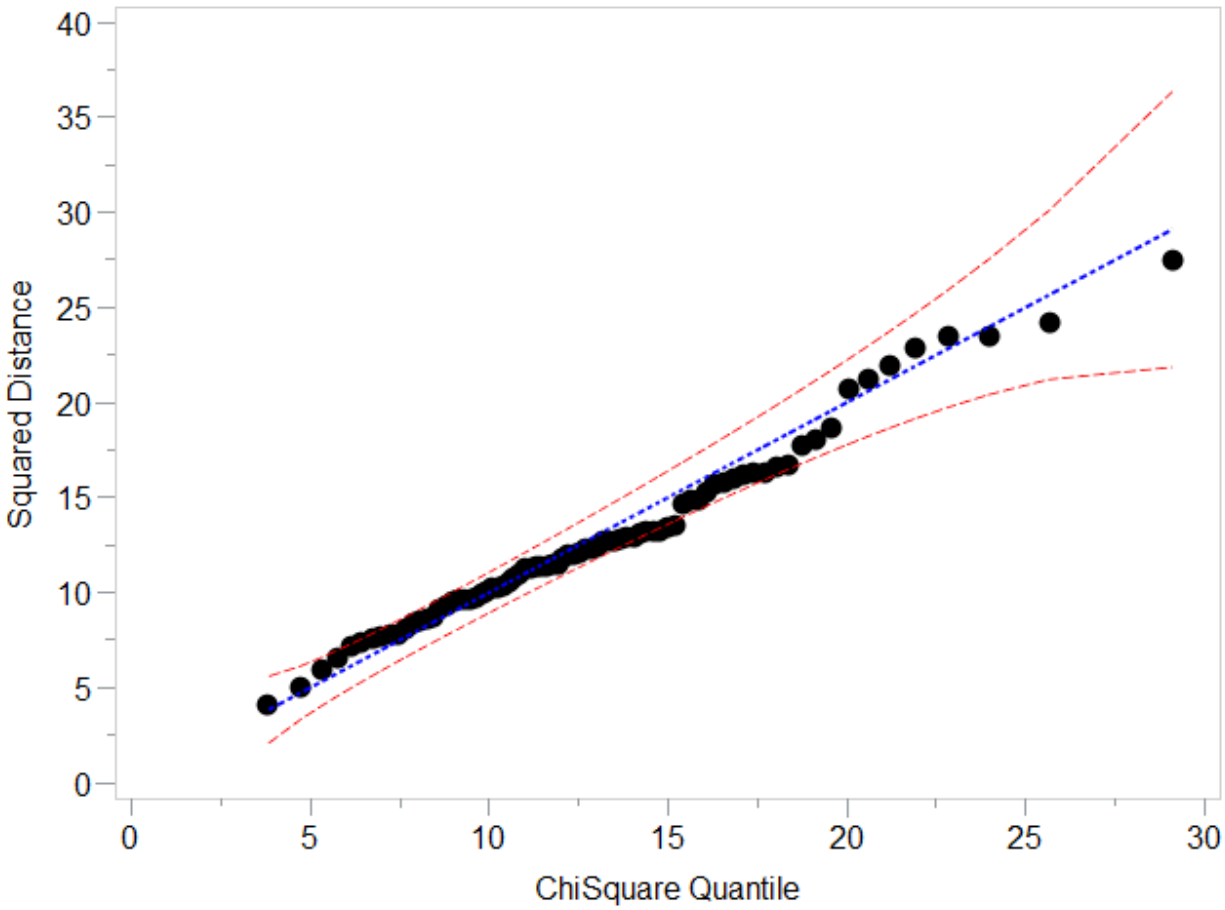

Figure 11. 13-variable LDA Black male c-q plot

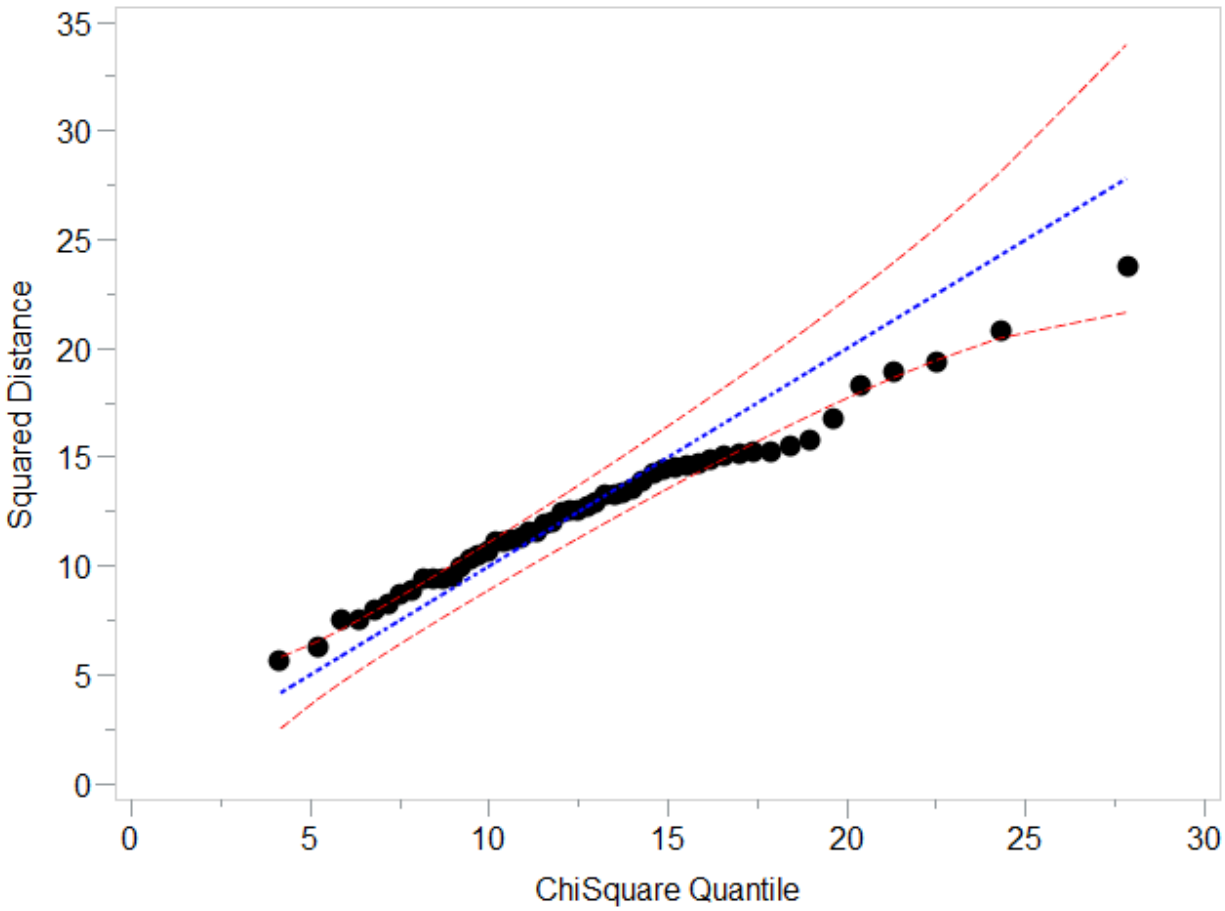

Figure 12. 13-variable LDA Black female c-q plot

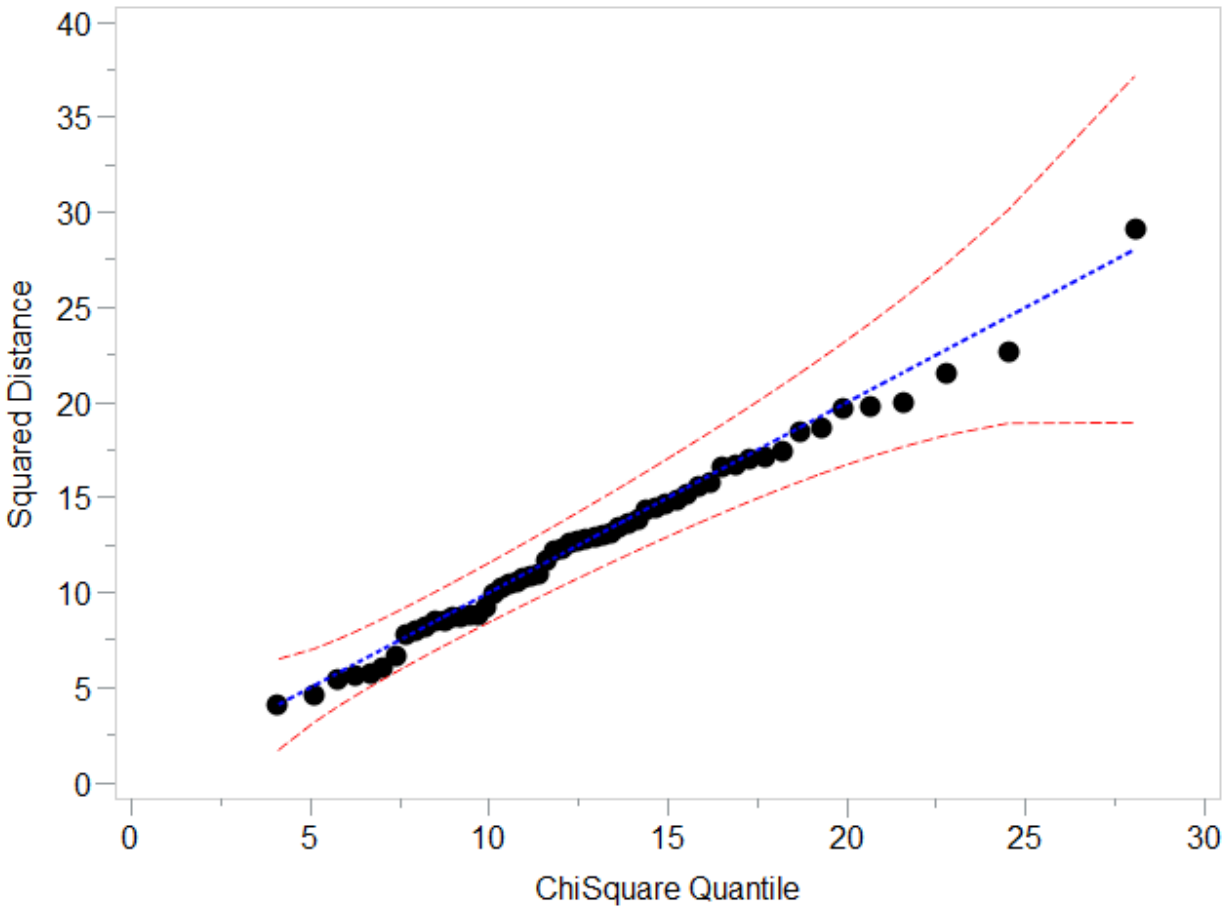

Figure 13. 13-variable LDA Hispanic male c-q plot

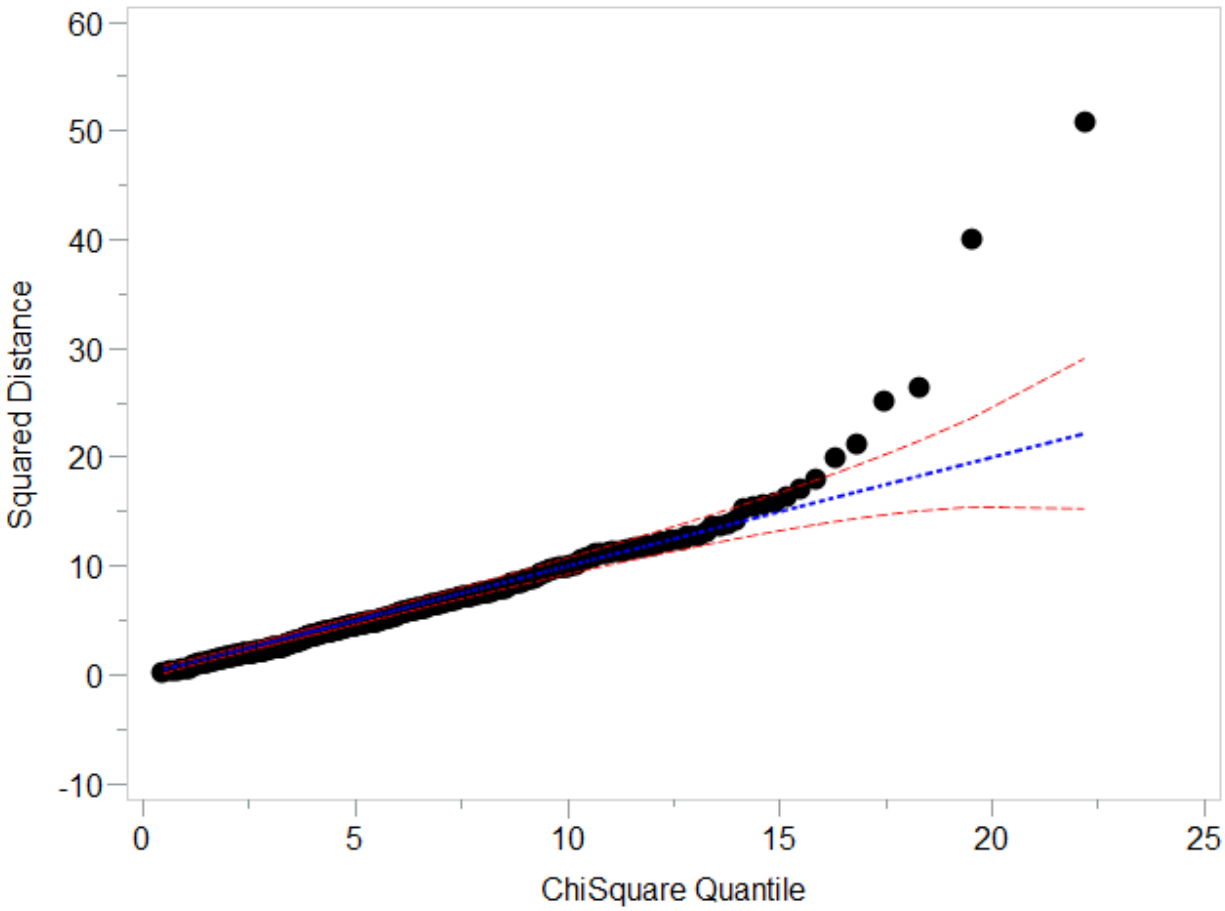

Figure 14. 6-variable LDA White male c-q plot

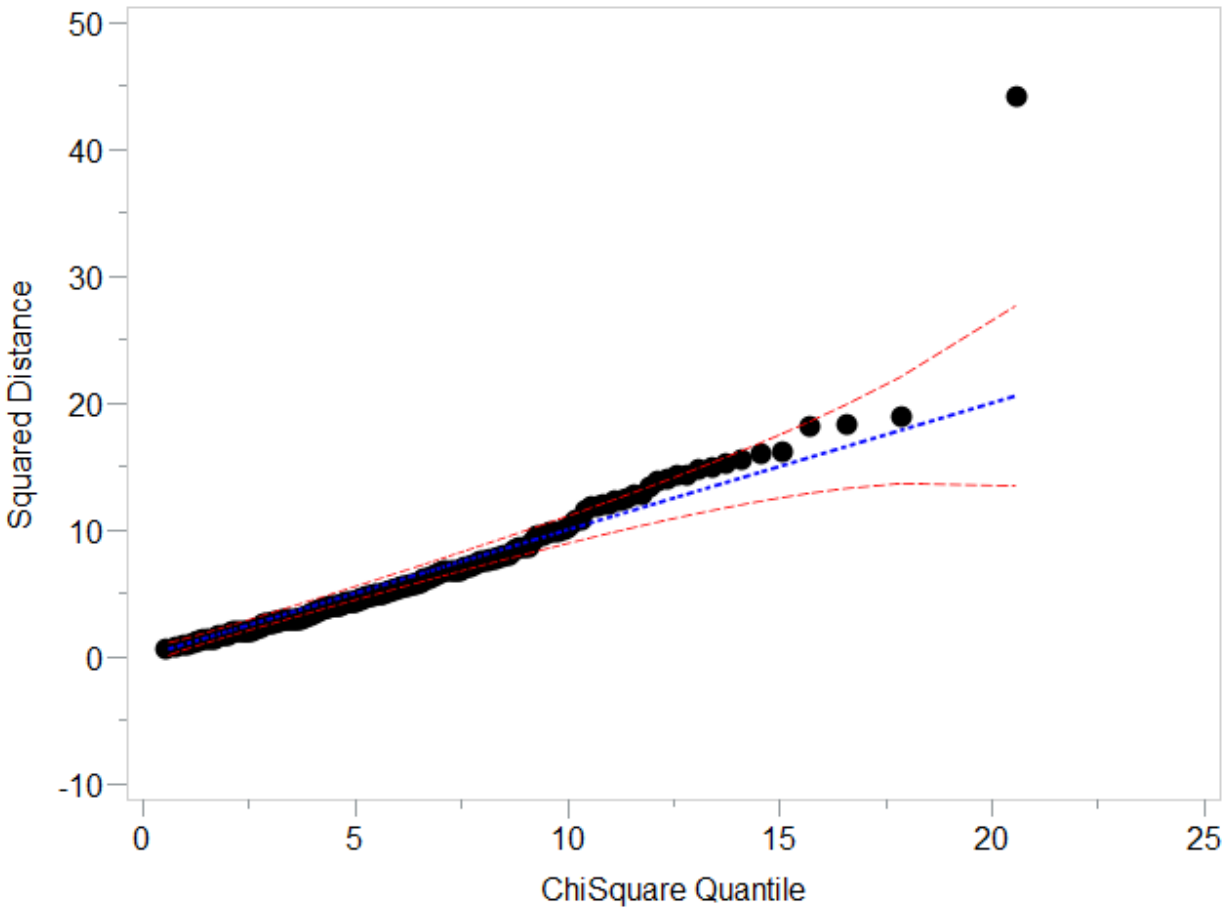

Figure 15. 6-variable LDA White female c-q plot

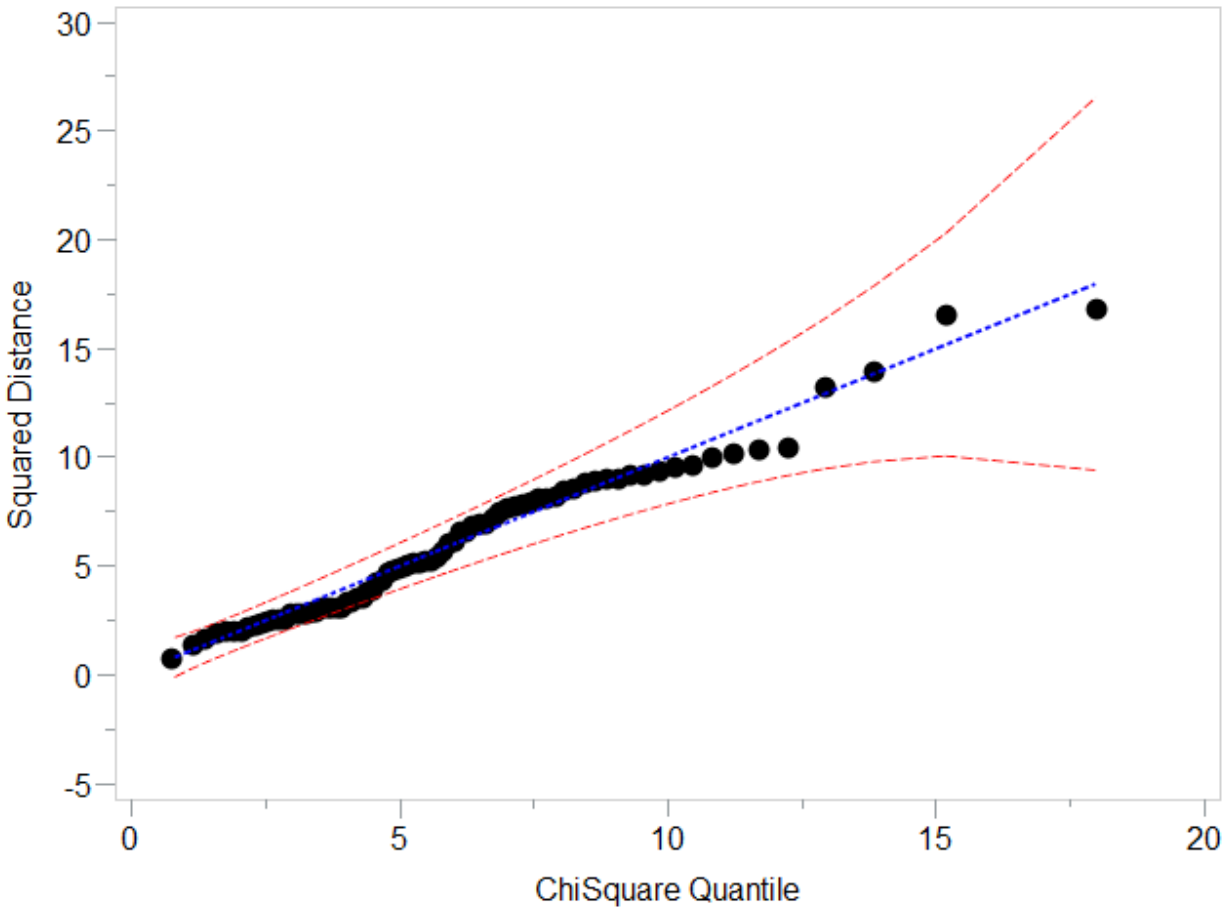

Figure 16. 6-variable LDA Black male c-q plot

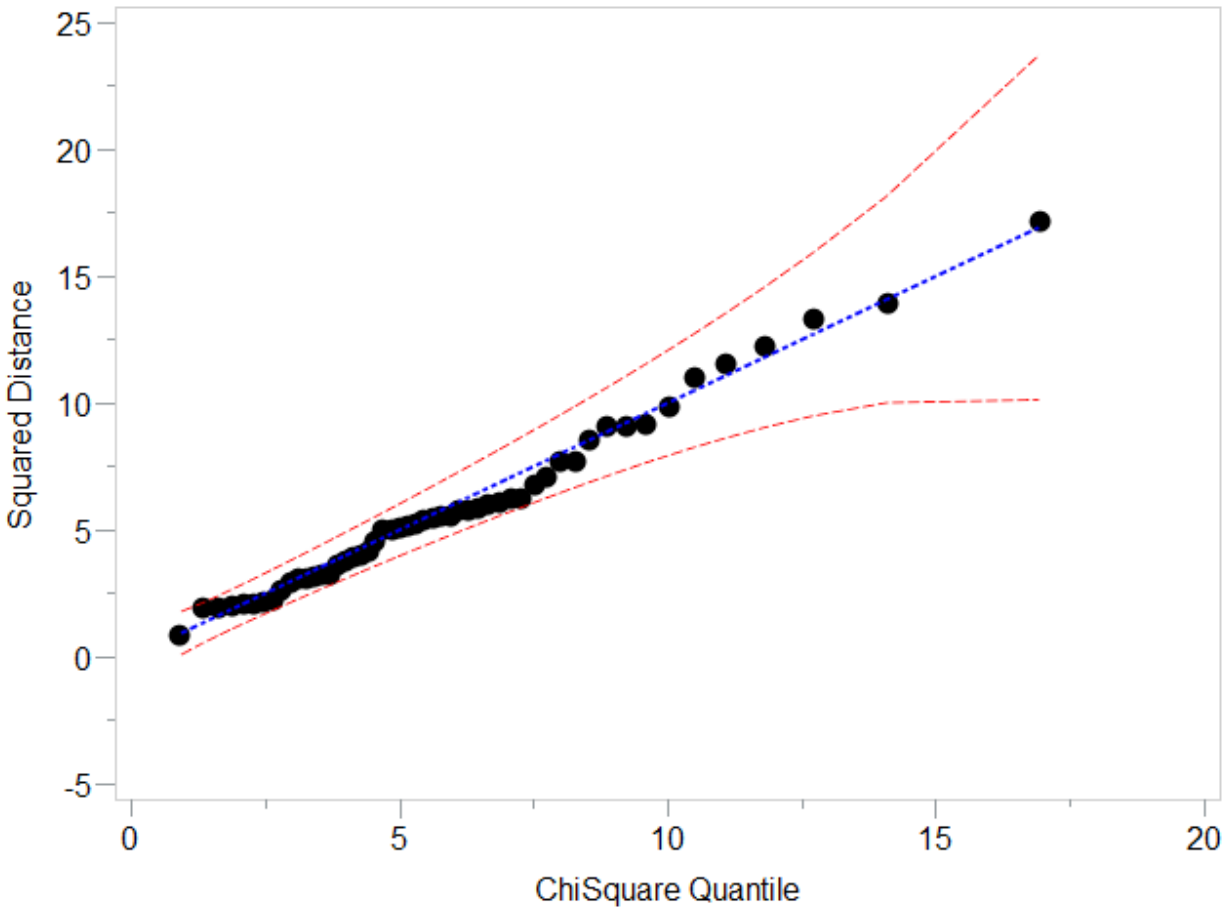

Figure 17. 6-variable LDA Black female c-q plot

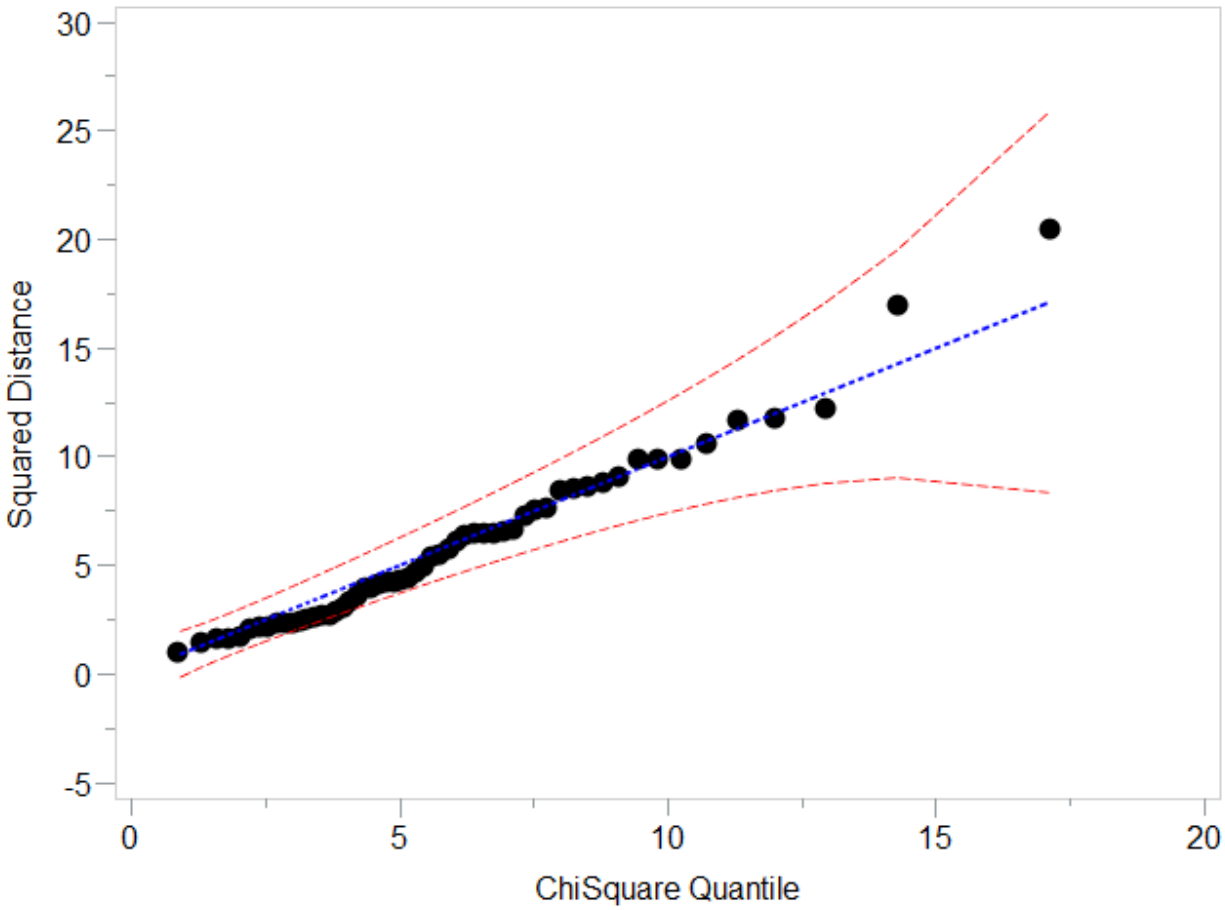

Figure 18. 6-variable LDA Hispanic male c-q plot

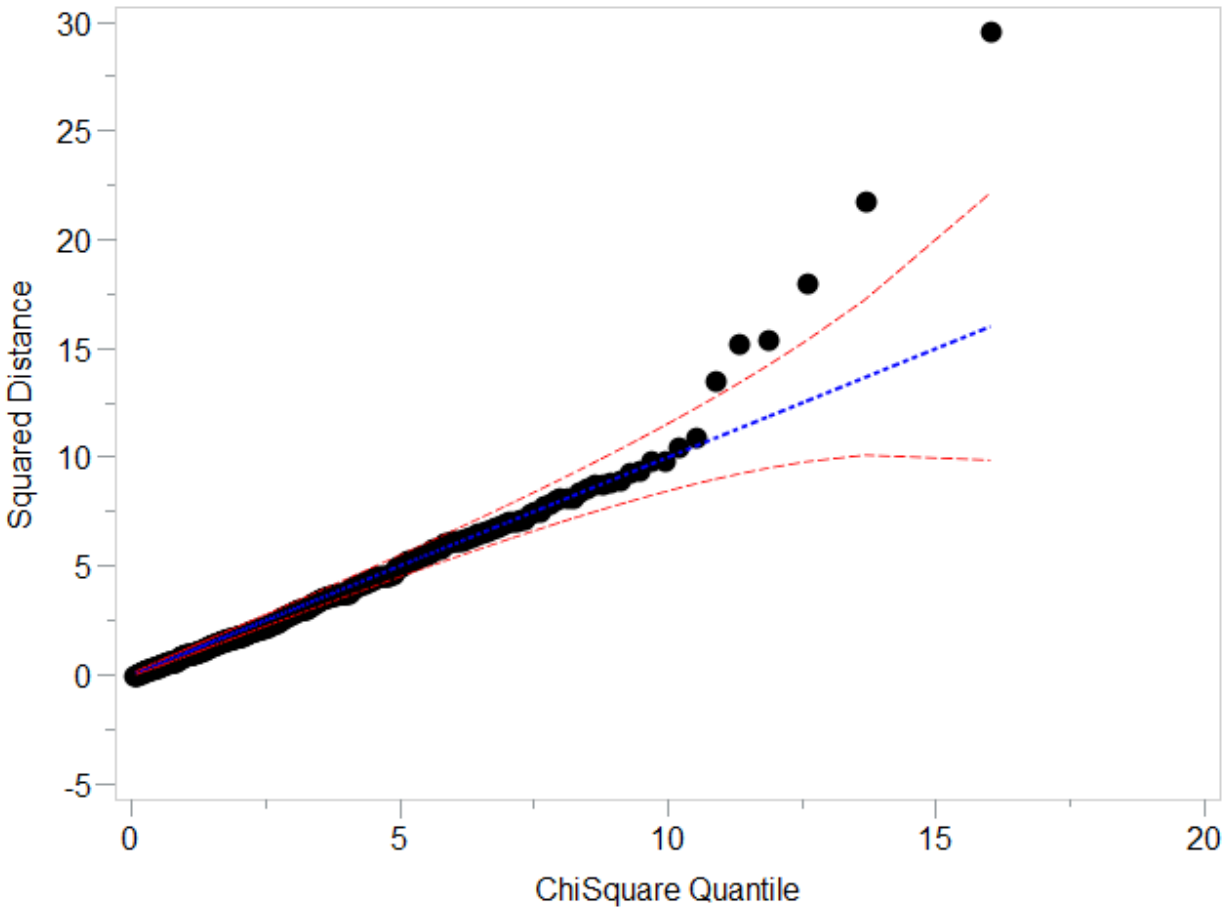

Figure 19. 3-variable QDA White male c-q plot

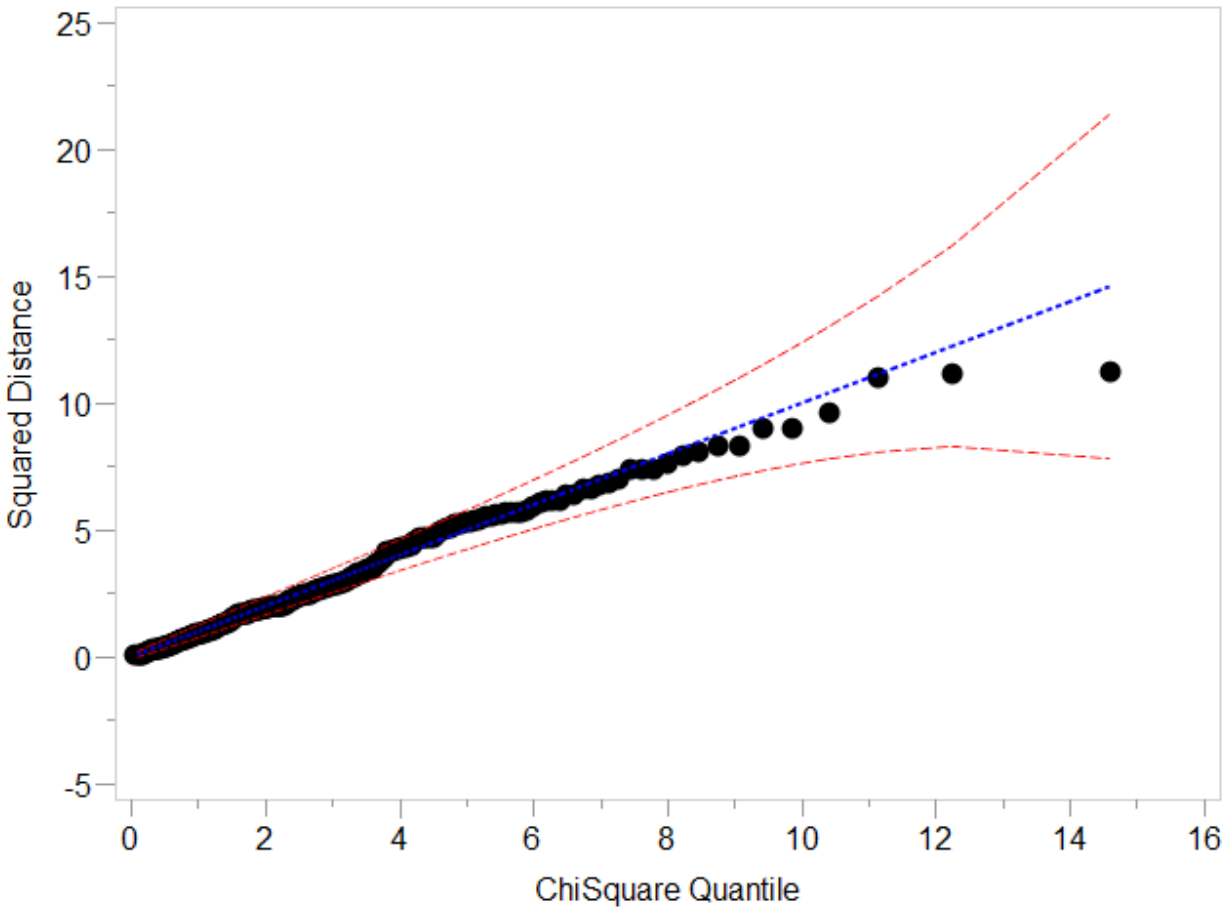

Figure 20. 3-variable QDA White female c-q plot

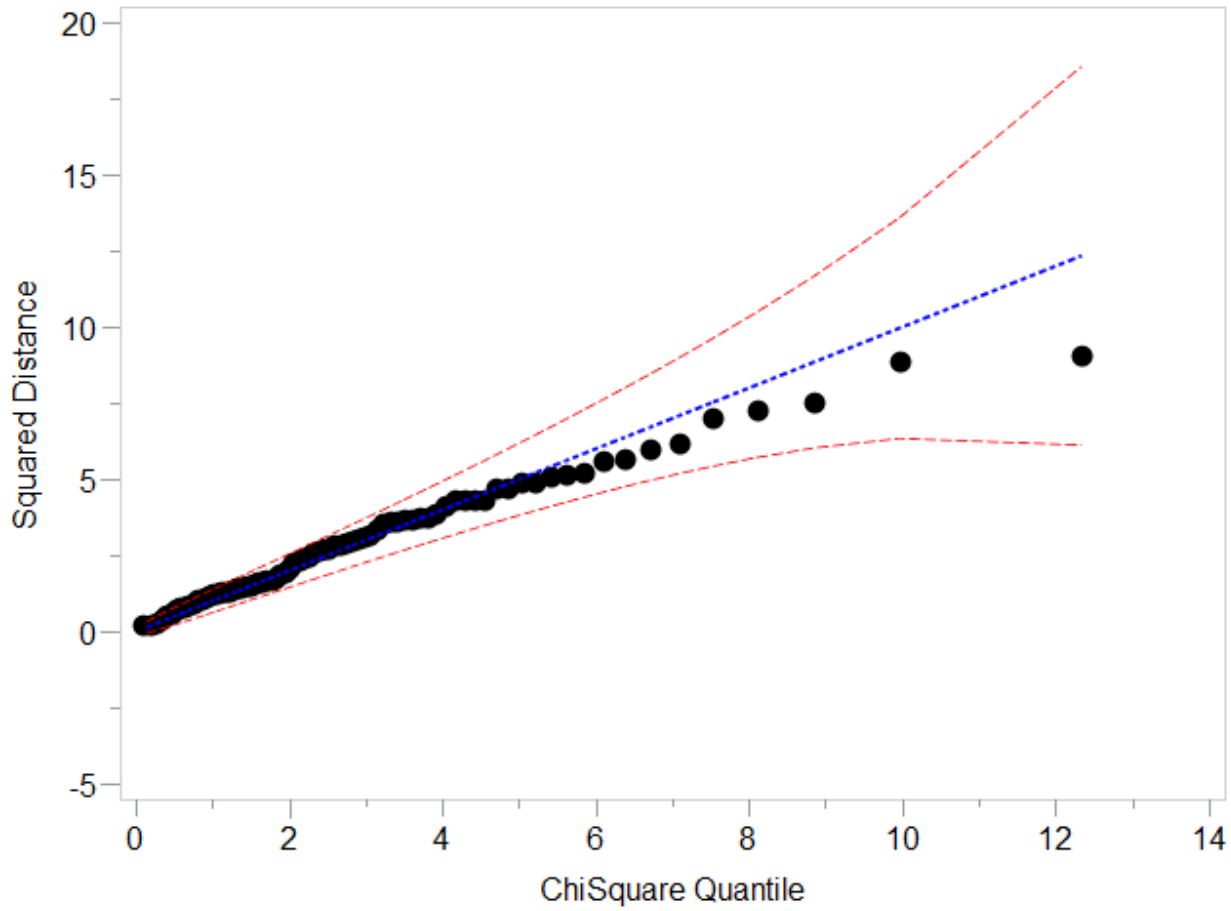

Figure 21. 3-variable QDA Black male c-q plot

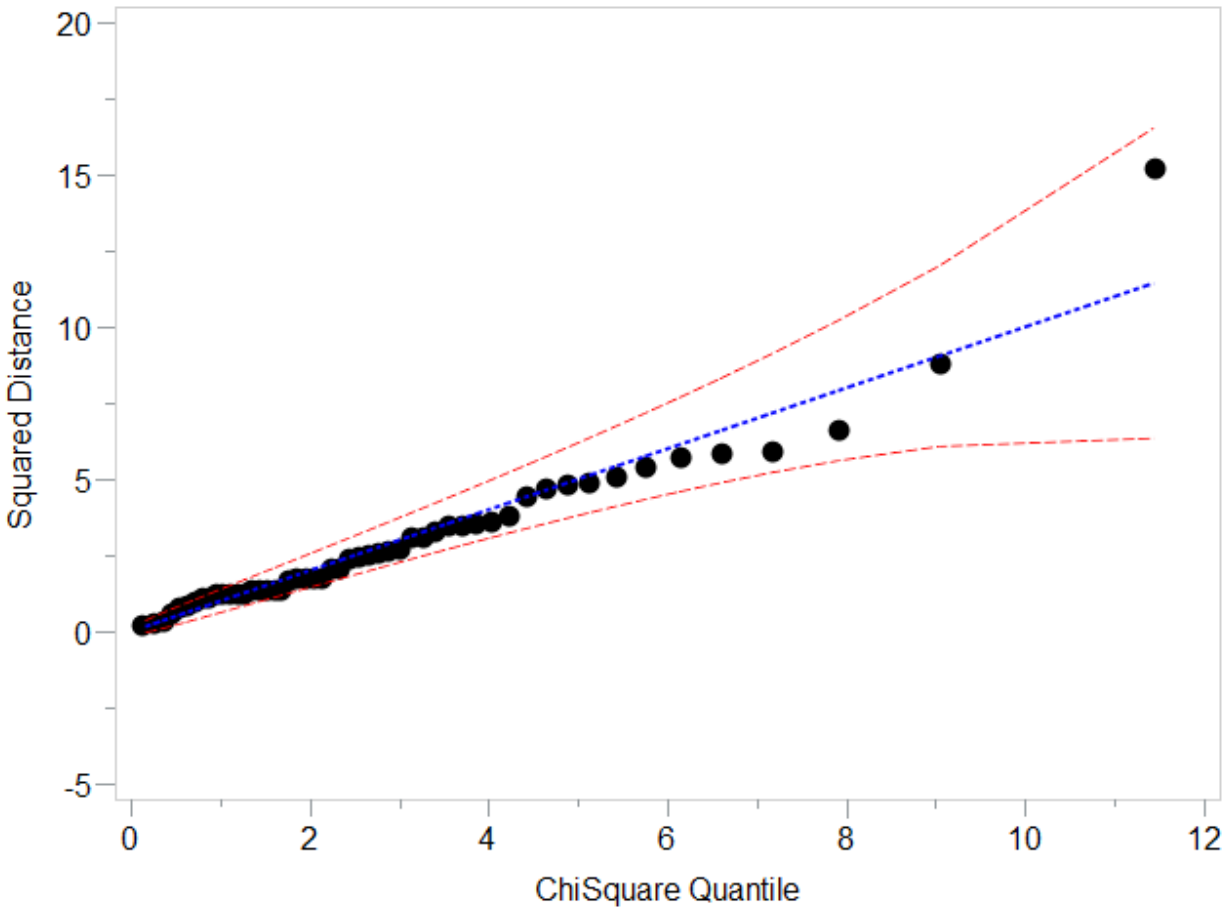

Figure 22. 3-variable QDA Black female c-q plot

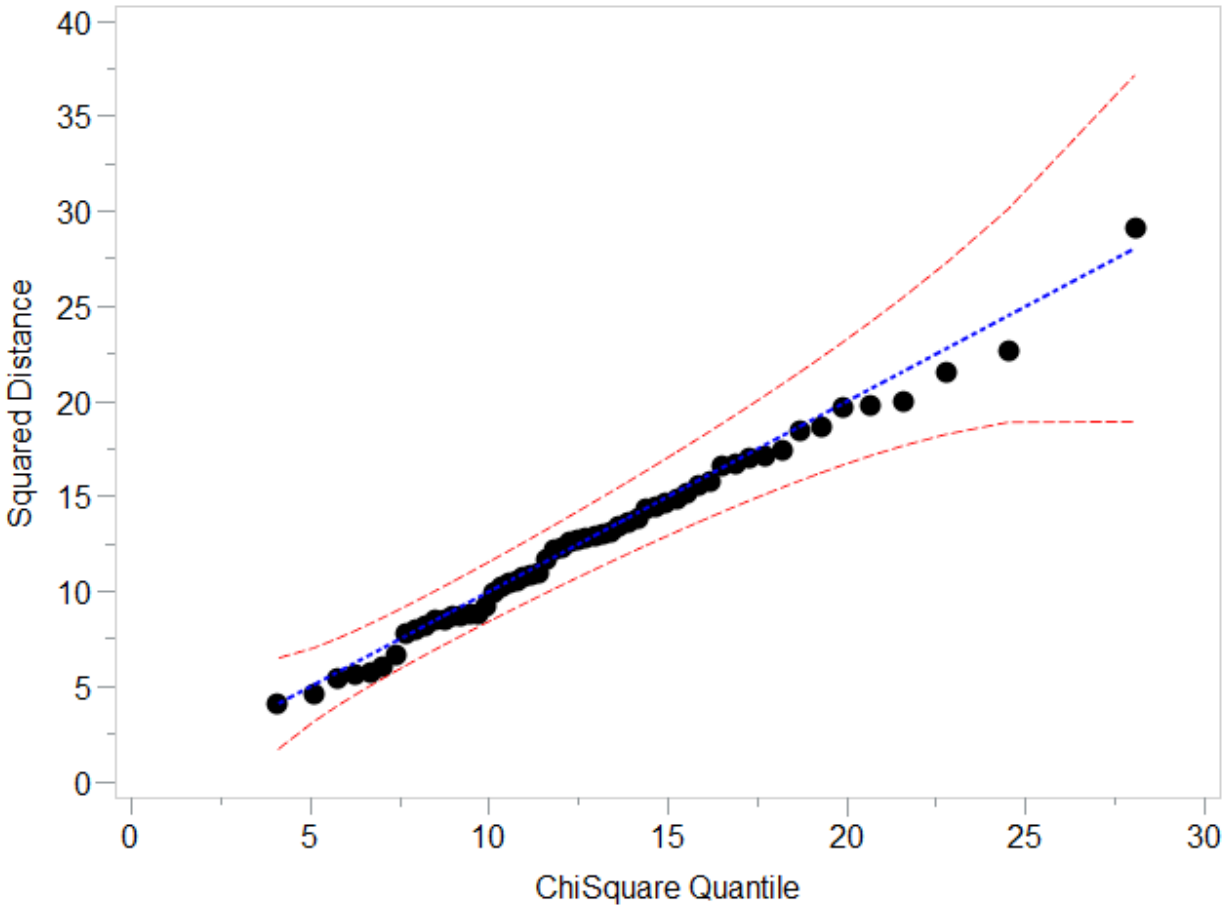

Figure 23. 3-variable QDA Hispanic male c-q plot

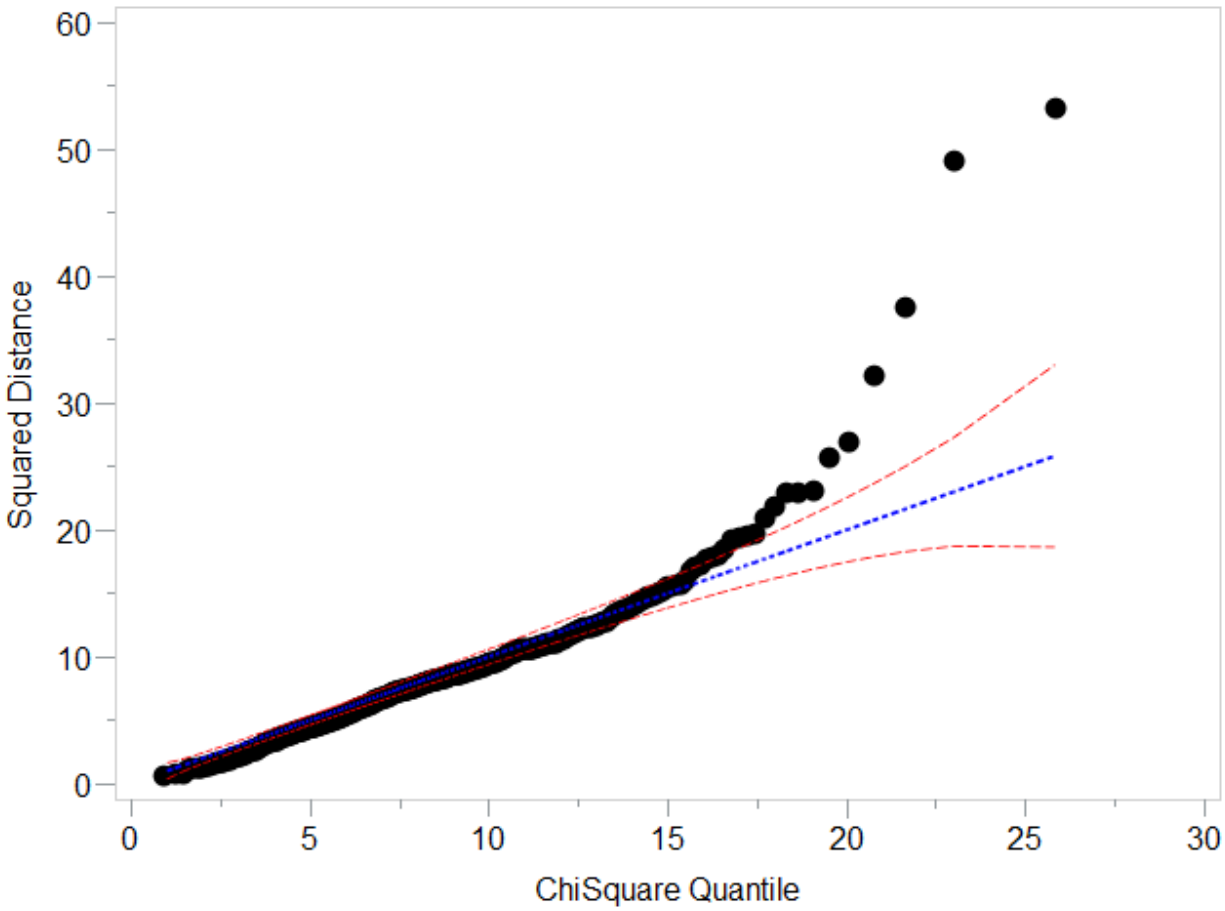

Figure 24. 8-variable KNN White male c-q plot

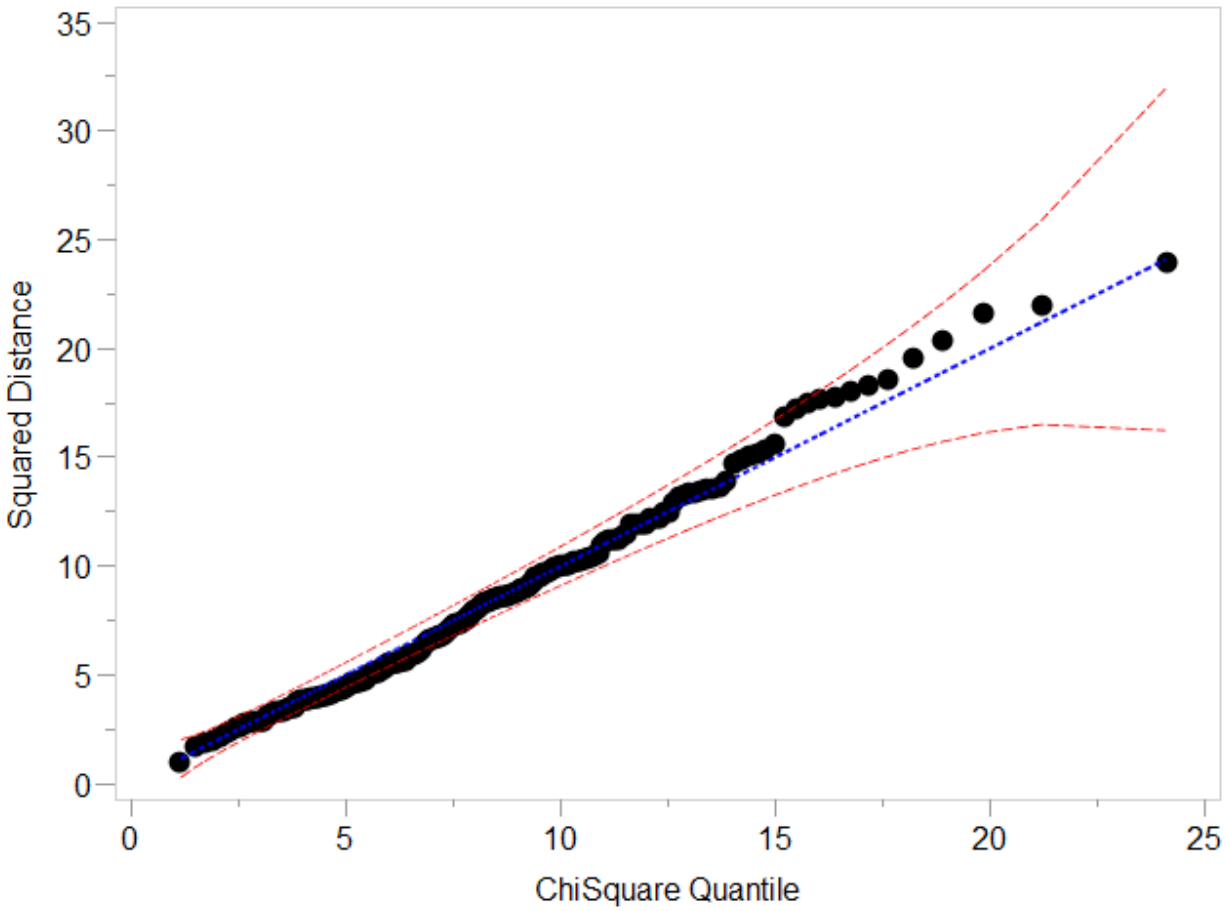

Figure 25. 8-variable KNN White female c-q plot

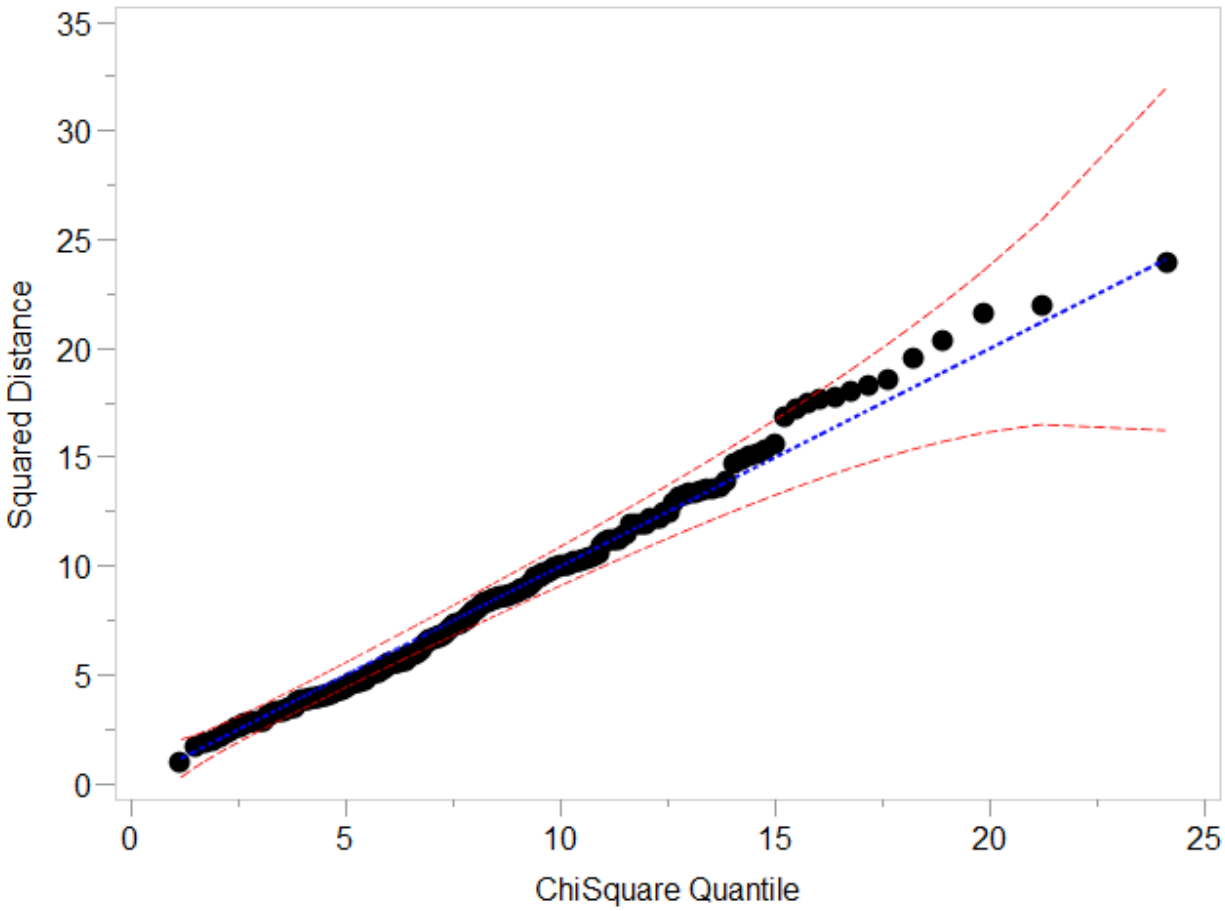

Figure 26. 8-variable KNN White female c-q plot

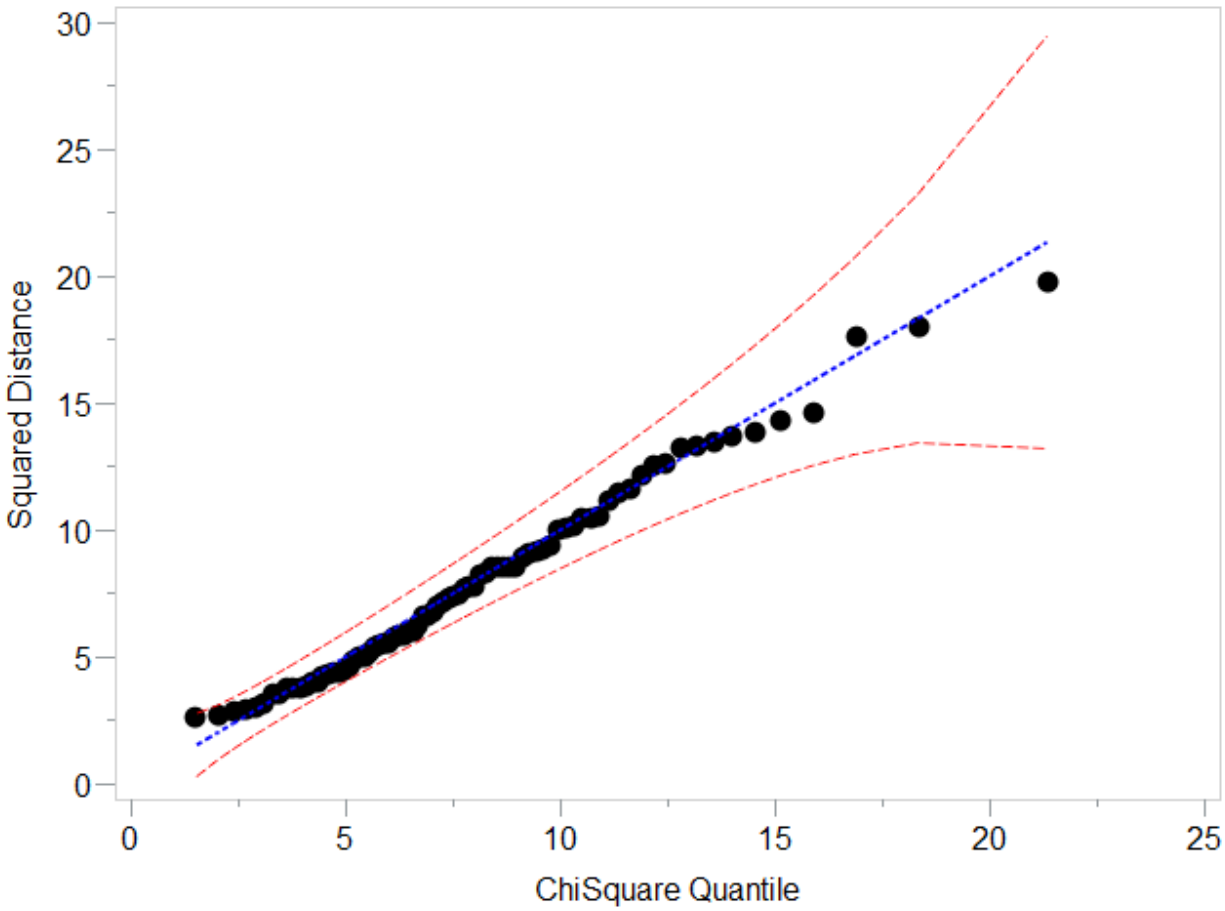

Figure 27. 8-variable KNN Black male c-q plot

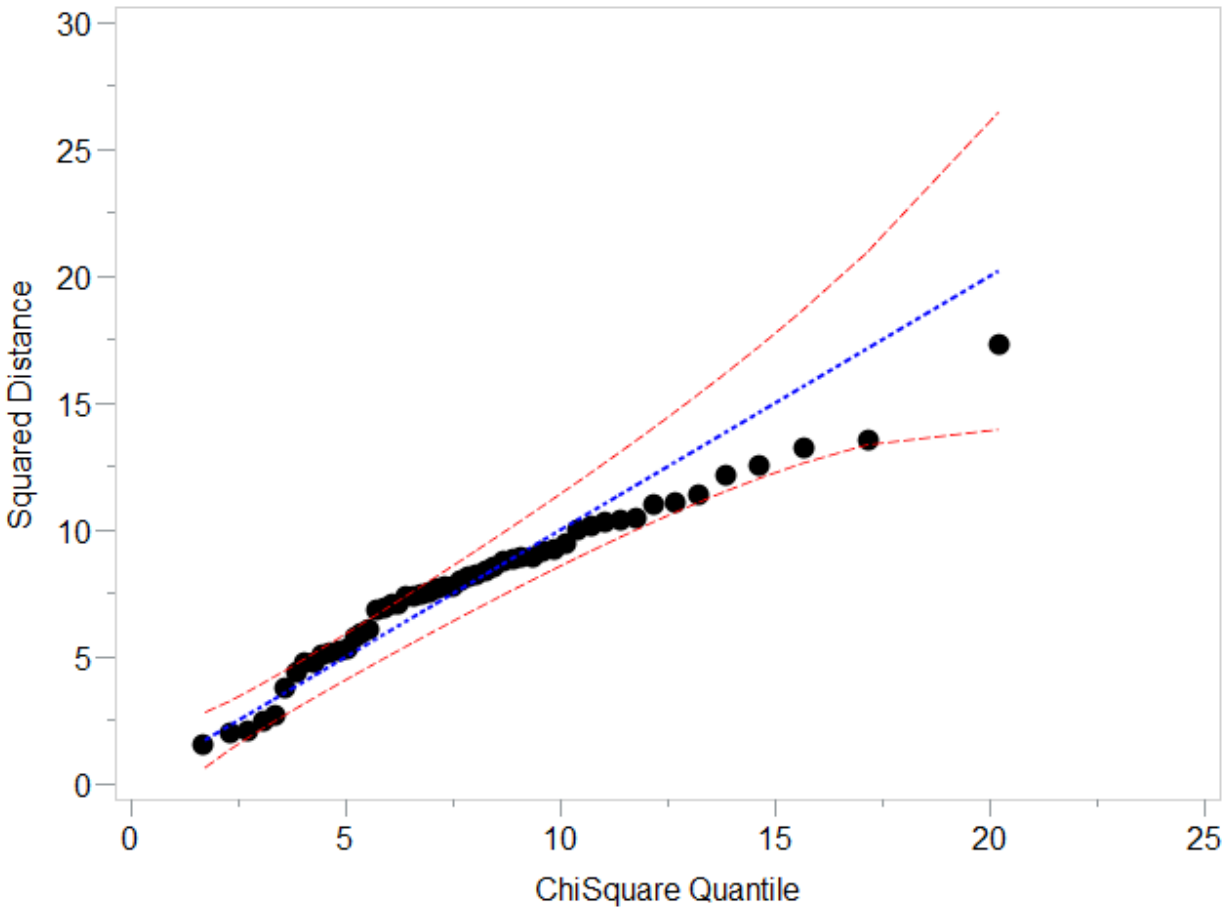

Figure 28. 8-variable KNN Black female c-q plot

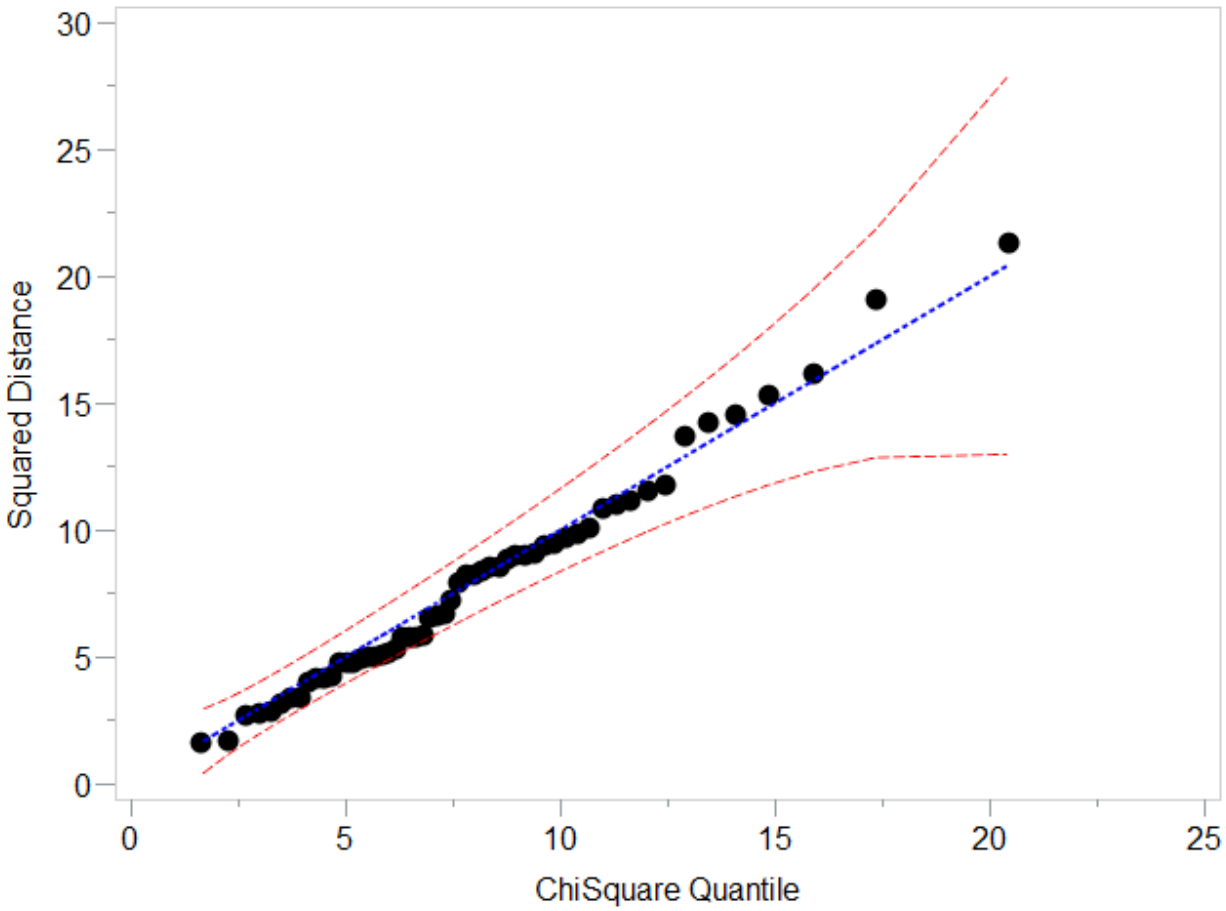

Figure 29. 8-variable KNN Hispanic male c-q plot

Table 22. 13 variable LDA ancestry aggregated cross-validation classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.7111	0.1781	0.1108	
B	0.1985	0.6641	0.1374	
H	0.1786	0.1786	0.6429	
Overall				0.6884

Table 23. 13 variable LDA ancestry aggregated holdout classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.774	0.137	0.089	
B	0.1803	0.6885	0.1311	
H	0.3182	0.1364	0.5455	
Overall				0.7547

Table 24. 6 variable LDA ancestry aggregated cross-validation classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.7111	0.1781	0.1108	
B	0.1985	0.6641	0.1374	
H	0.1786	0.1786	0.6429	
Overall				0.6994

Table 25. 6 variable LDA ancestry aggregated holdout classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.7397	0.1712	0.1473	
B	0.1475	0.7049	0.1475	
H	0.2727	0.1818	0.5455	
Overall				0.7227

Table 26. 3 variable QDA ancestry aggregated cross-validation classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.6287	0.2006	0.1707	
B	0.2443	0.6183	0.1374	
H	0.2857	0.1964	0.5179	
Overall				0.6199

Table 27. 3 variable QDA ancestry aggregated holdout classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.6644	0.1884	0.1473	
B	0.1967	0.623	0.1803	
H	0.4091	0.0909	0.5	
Overall				0.648

Table 28. 6 variable QDA ancestry aggregated cross-validation classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.7066	0.1722	0.1213	
B	0.229	0.687	0.084	
H	0.2679	0.1607	0.5926	
Overall				0.6947

Table 29. 6 variable QDA ancestry aggregated holdout classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.7123	0.1952	0.0925	
B	0.2131	0.6557	0.1311	
H	0.3182	0.1363	0.5455	
Overall				0.6933

Table 30. 5 variable KNN ancestry aggregated cross-validation classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.6048	0.274	0.1212	
B	0.3282	0.5802	0.0916	
H	0.4107	0.375	0.2142	
Overall				0.5754

Table 31. 5 variable KNN ancestry aggregated holdout classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.6336	0.2842	0.0822	
B	0.2459	0.5902	0.1639	
H	0.45	0.05	0.5	
Overall				0.6187

Table 32. 8 variable KNN ancestry aggregated cross-validation classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.6766	0.2171	0.1062	
B	0.2824	0.5649	0.1527	
H	0.4464	0.2857	0.2679	
Overall				0.5895

Table 33. 8 variable KNN ancestry aggregated holdout classification rate

	Predicted			
Actual	**W**	**B**	**H**	**Overall**
W	0.6815	0.2021	0.0822	
B	0.2951	0.541	0.1639	
H	0.3636	0.0909	0.5455	
Overall				0.6507

Table 34. 13 variable LDA sex aggregated cross-validation classification rate

	Predicted		
Actual	**M**	**F**	**Overall**
M	0.9255	0.0745	
F	0.1906	0.8094	
Overall			0.8877

Table 35. 13 variable LDA sex aggregated holdout classification rate

	Predicted		
Actual	**M**	**F**	**Overall**
M	0.8988	0.1012	
F	0.1172	0.8828	
Overall			0.8933

Table 36. 6 variable LDA sex aggregated cross-validation classification rate

	Predicted		
Actual	**M**	**F**	**Overall**
M	0.9255	0.0745	
F	0.1906	0.8094	
Overall			0.8877

Table 37. 6 variable LDA sex aggregated holdout classification rate

	Predicted		
Actual	**M**	**F**	**Overall**
M	0.9028	0.0972	
F	0.1562	0.8438	
Overall			0.8827

Table 38. 3 variable QDA sex aggregated cross-validation classification rate

	Predicted		
Actual	**M**	**F**	**Overall**
M	0.825	0.175	
F	0.2842	0.7158	
Overall			0.7895

Table 39. 3 variable QDA sex aggregated holdout classification rate

	Predicted		
Actual	**M**	**F**	**Overall**
M	0.8057	0.1943	
F	0.2578	0.7422	
Overall			0.784

Table 40. 6 variable QDA sex aggregated cross-validation classification rate

	Predicted		
Actual	M	F	Overall
M	0.8925	0.1075	
F	0.2662	0.7338	
Overall			0.8409

Table 41. 6 variable QDA sex aggregated holdout classification rate

	Predicted		
Actual	**M**	**F**	**Overall**
M	0.8745	0.1255	
F	0.2891	0.7109	
Overall			0.8187

Table 42. 5 variable KNN sex aggregated cross-validation classification rate

	Predicted		
Actual	**M**	**F**	**Overall**
M	0.8683	0.1317	
F	0.2698	0.7302	
Overall			0.8234

Table 43. 5 variable KNN sex aggregated holdout classification rate

	Predicted		
Actual	**M**	**F**	**Overall**
M	0.8704	0.1296	
F	0.2031	0.7969	
Overall			0.8453

Table 44. 8 variable KNN sex aggregated cross-validation classification rate

	Predicted		
Actual	**M**	**F**	**Overall**
M	0.8719	0.1281	
F	0.2986	0.7014	
Overall			0.8164

Table 45. 8 variable KNN sex aggregated holdout classification rate

	Predicted		
Actual	M	F	Overall
M	0.8623	0.1377	
F	0.2891	0.7109	
Overall			0.8107

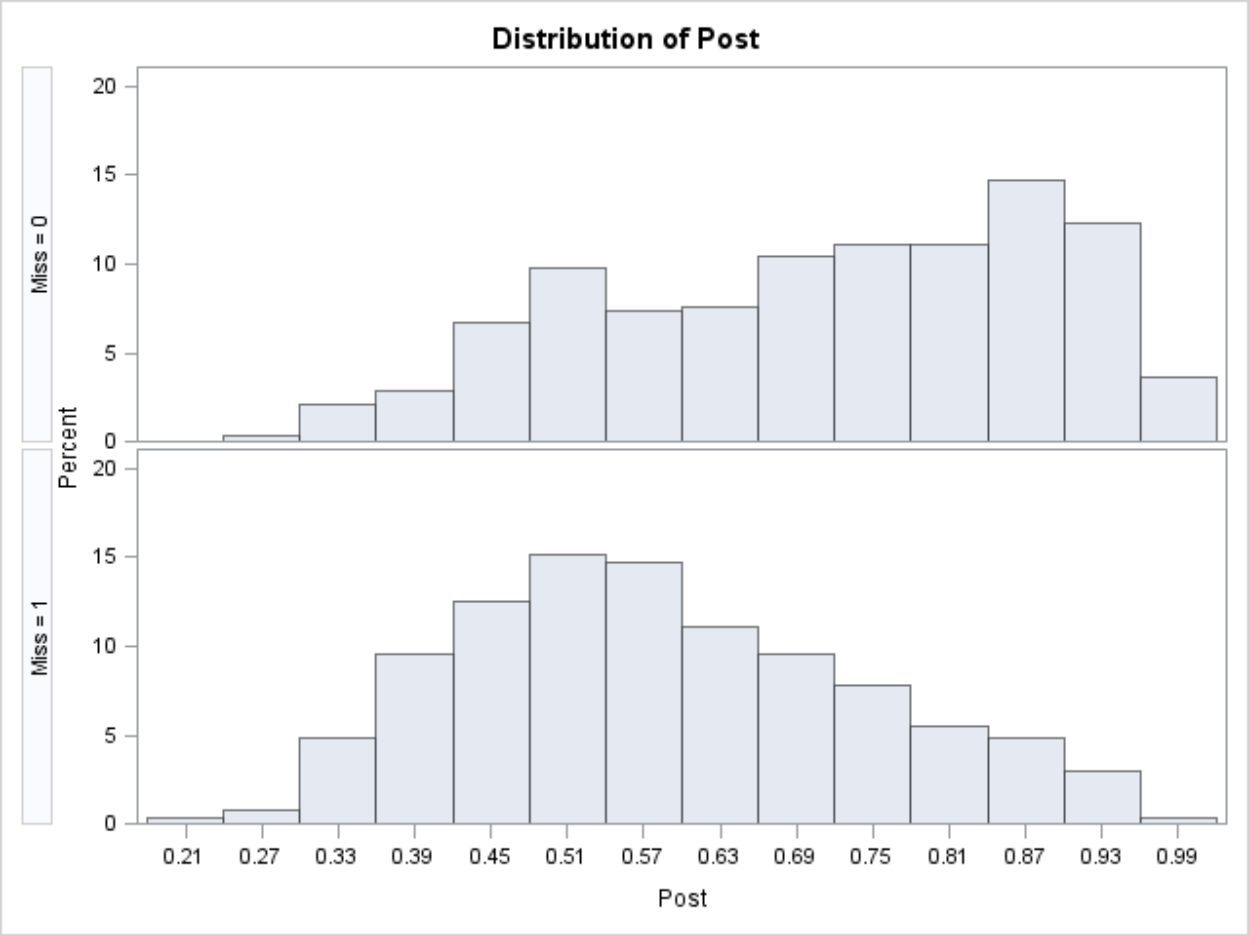

Figure 30. 13-variable LDA cross-validation posterior probabilities histogram

Basic Statistical Measures			
Location		Variability	
Mean	0.584649	Std Deviation	0.15821
Median	0.566354	Variance	0.02503
Mode	.	Range	0.75266
		Interquartile Range	0.22196

Figure 31. Descriptive statistics - LDA cross-validation misclassification posterior probabilities

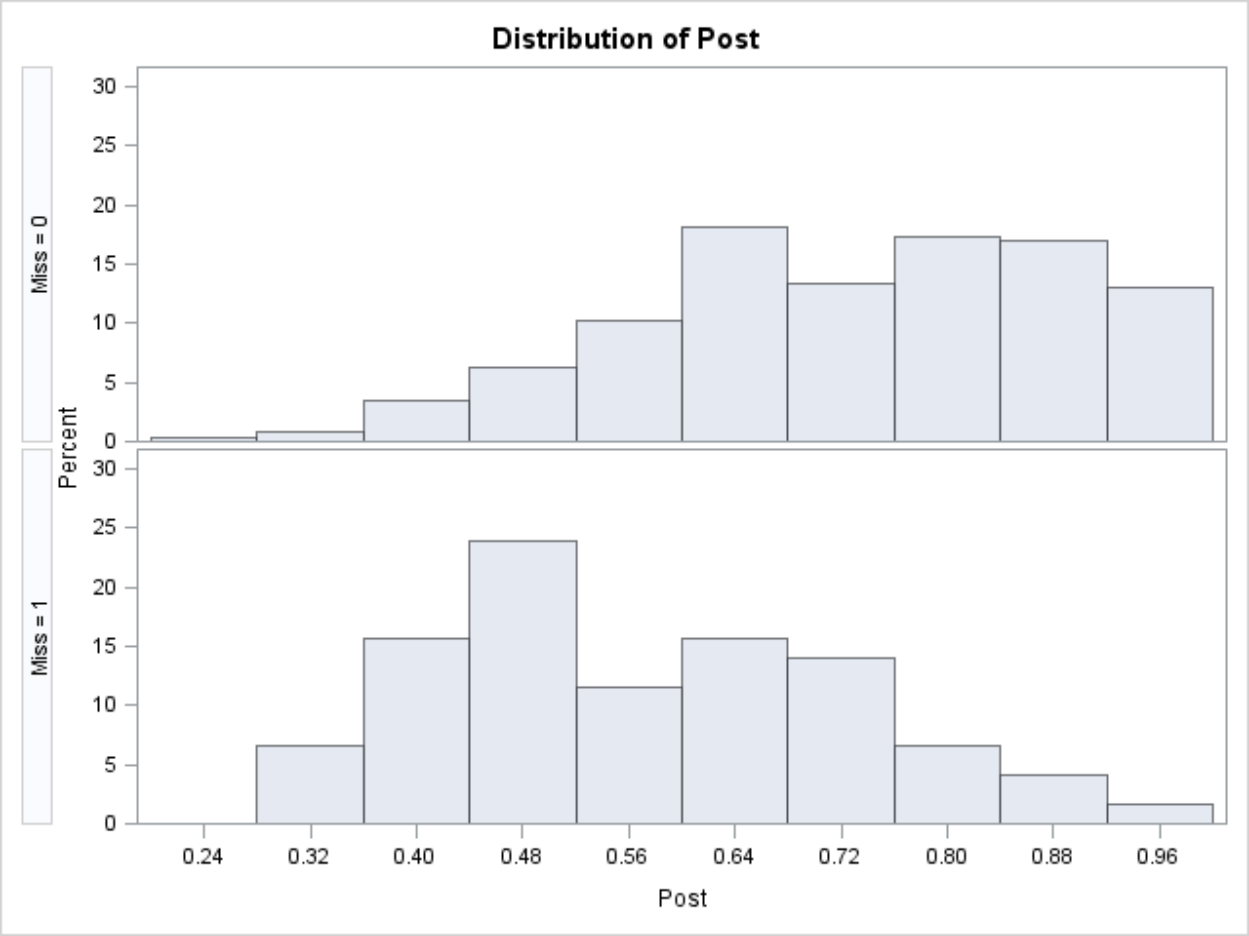

Figure 32. 13-variable LDA holdout posterior probabilities histogram

Basic Statistical Measures			
Location		Variability	
Mean	0.568582	Std Deviation	0.15643
Median	0.542723	Variance	0.02447
Mode	0.494414	Range	0.67187
		Interquartile Range	0.23676

Figure 33. Descriptive statistics- LDA holdout misclassification posterior probabilities

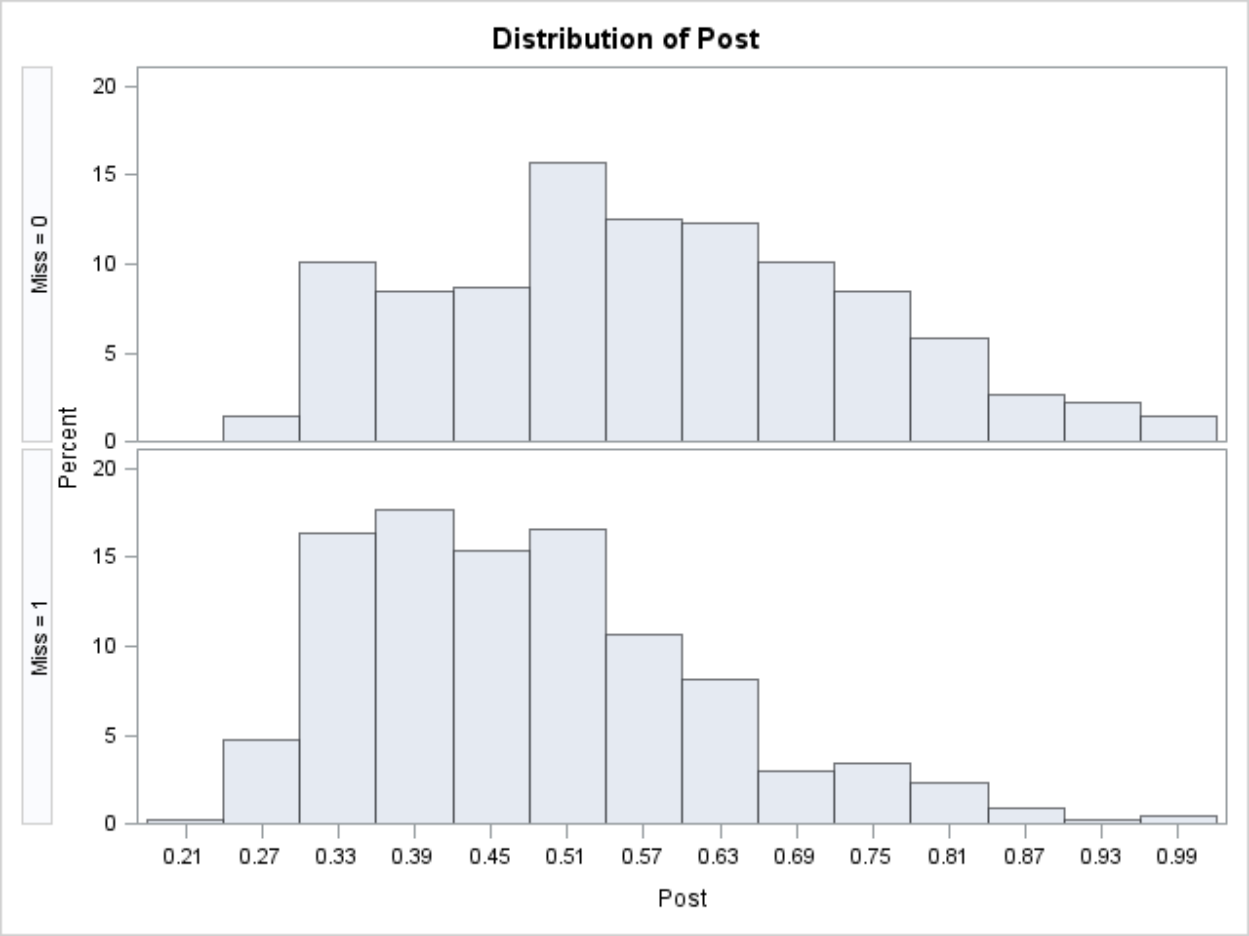

Figure 34. 3-variable QDA cross-validation posterior probabilities histogram

Basic Statistical Measures			
Location		Variability	
Mean	0.480560	Std Deviation	0.13941
Median	0.459729	Variance	0.01944
Mode	0.324499	Range	0.76018
		Interquartile Range	0.19223

Figure 35. Descriptive statistics - QDA cross-validation misclassification posterior probabilites

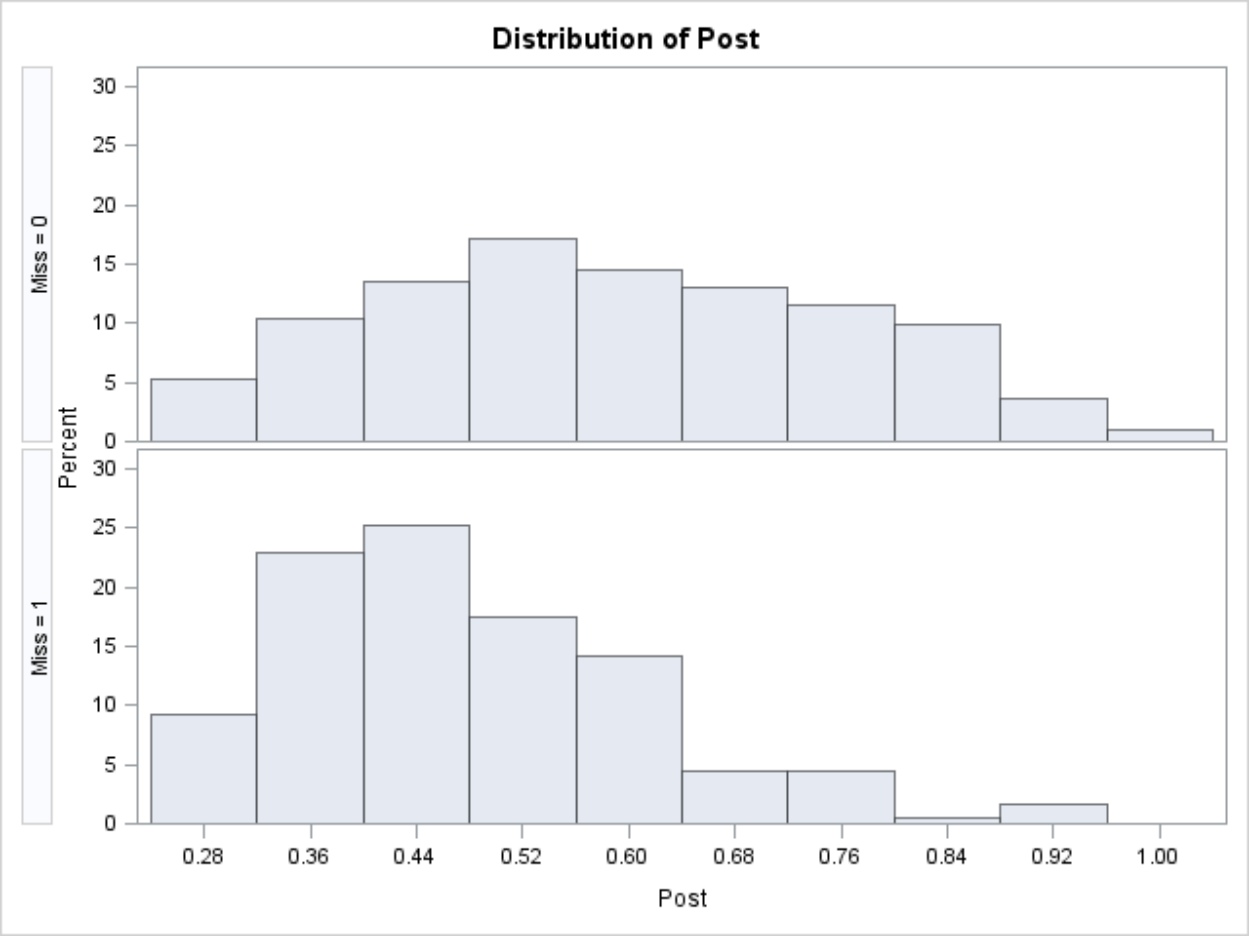

Figure 36. 3-variable QDA holdout posterior probabilities histogram

Basic Statistical Measures			
Location		Variability	
Mean	0.477831	Std Deviation	0.13788
Median	0.447222	Variance	0.01901
Mode	0.282362	Range	0.69865
		Interquartile Range	0.19297

Figure 37. Descriptive statistics - QDA holdout misclassification posterior probabilities

Basic Statistical Measures			
Location		Variability	
Mean	0.353074	Std Deviation	0.24699
Median	0.322565	Variance	0.06100
Mode	.	Range	0.98248
		Interquartile Range	0.38265

Figure 38. Descriptive statistics - LDA re-substitution misses

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	0.982781516
99%	0.904688248
95%	0.847117106
90%	0.721370664
75% Q3	0.520794806
50% Median	0.322565207
25% Q1	0.138141773
10%	0.048522546
5%	0.024050301
1%	0.007213173
0% Min	0.000304265

Figure 39. LDA re-substitution misses quartile statistics

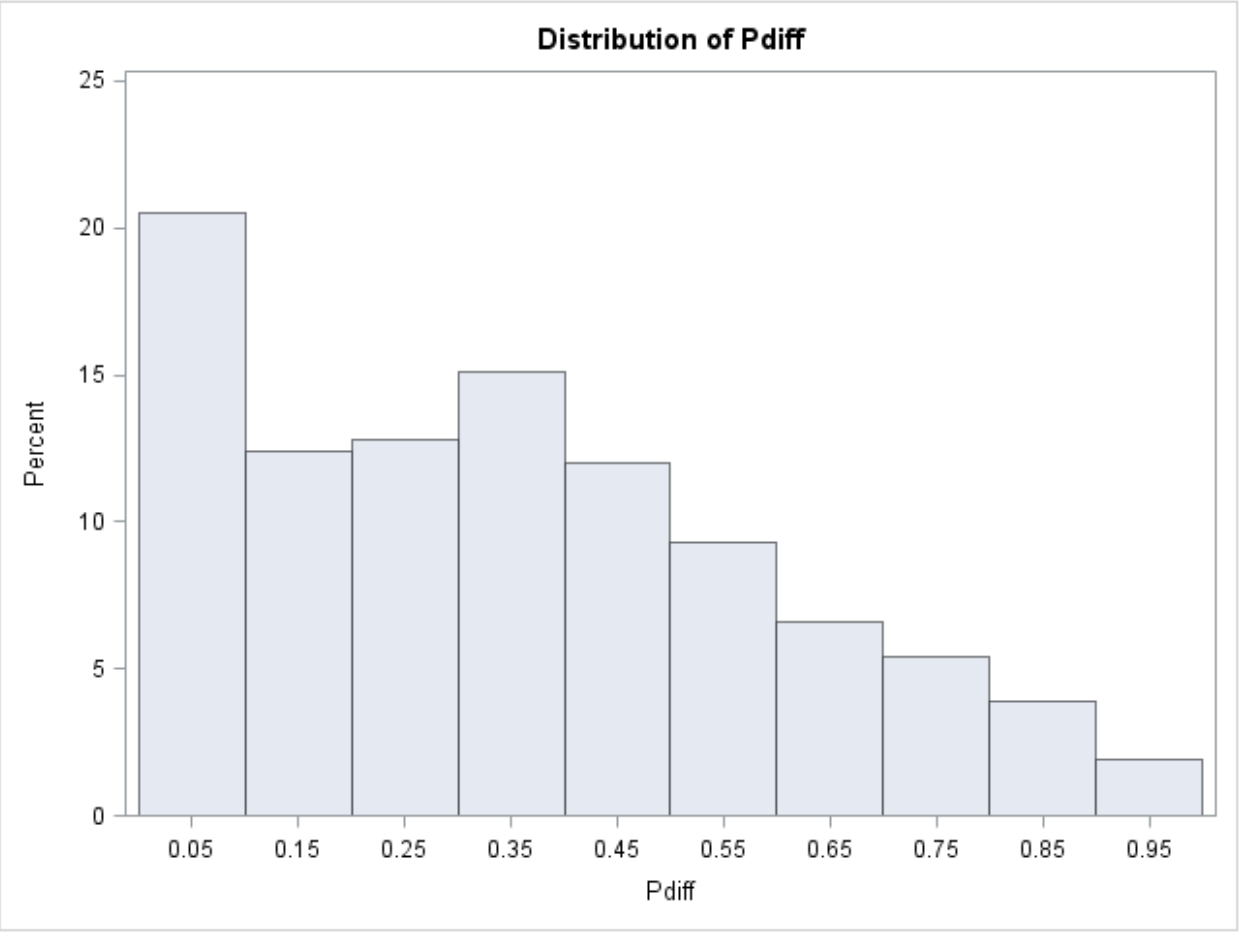

Figure 40. LDA re-substitution misses histogram

Basic Statistical Measures			
Location		Variability	
Mean	0.369009	Std Deviation	0.24629
Median	0.343341	Variance	0.06066
Mode	.	Range	0.97864
		Interquartile Range	0.39066

Figure 41. Descriptive statistics - LDA cross-validation misses

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	0.98508089
99%	0.92109740
95%	0.84647592
90%	0.73933758
75% Q3	0.53408718
50% Median	0.34334086
25% Q1	0.14342380
10%	0.06429676
5%	0.03678127
1%	0.01290807
0% Min	0.00644165

Figure 42. LDA cross-validation misses quartiles

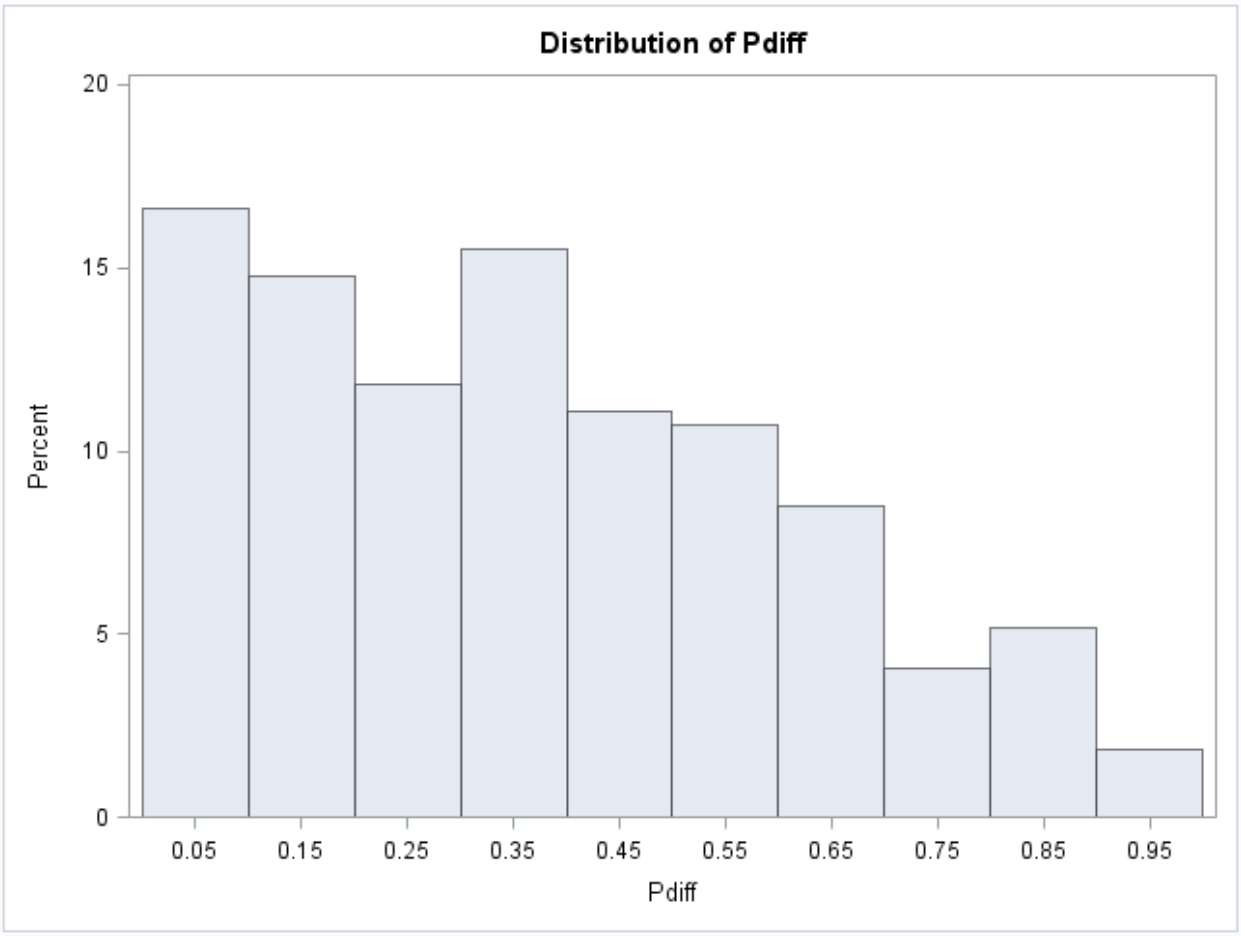

Figure 43. LDA cross-validation misses histogram

Basic Statistical Measures			
Location		Variability	
Mean	0.348359	Std Deviation	0.24304
Median	0.337470	Variance	0.05907
Mode	0.342624	Range	0.95639
		Interquartile Range	0.37542

Figure 44. Descriptive statistics - LDA holdout misses

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	0.95863718
99%	0.93427360
95%	0.80083376
90%	0.65893912
75% Q3	0.53029838
50% Median	0.33746968
25% Q1	0.15488213
10%	0.05175630
5%	0.02361759
1%	0.00402274
0% Min	0.00224369

Figure 45. LDA holdout misses quartiles

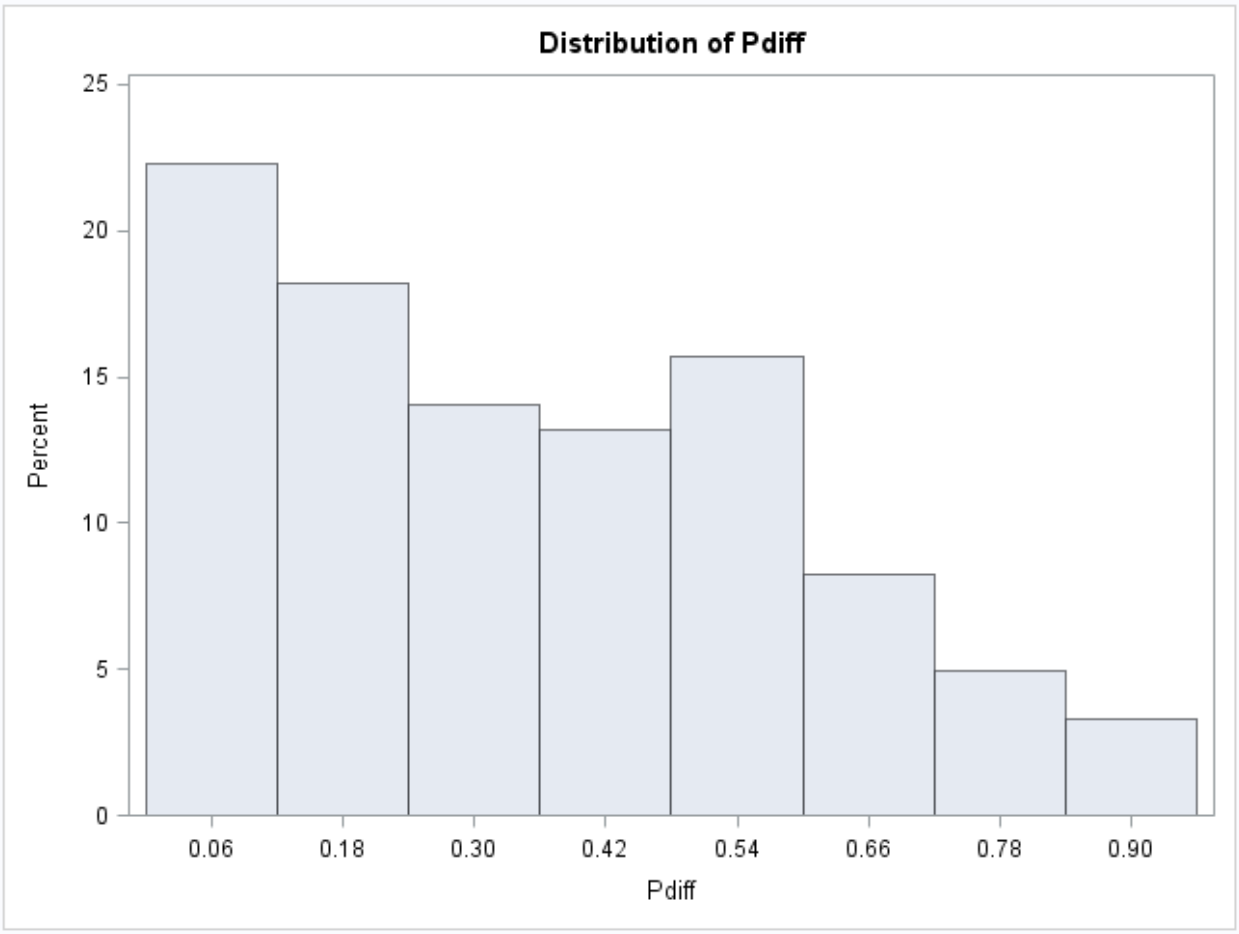

Figure 46. LDA holdout misses histogram

Basic Statistical Measures			
Location		Variability	
Mean	0.272729	Std Deviation	0.18615
Median	0.248689	Variance	0.03465
Mode	0.076765	Range	0.99611
		Interquartile Range	0.25227

Figure 47. Descriptive statistics - QDA resubstitution misses

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	0.997012413
99%	0.788936930
95%	0.629097643
90%	0.538149197
75% Q3	0.382616622
50% Median	0.248688957
25% Q1	0.130342845
10%	0.050814752
5%	0.023065133
1%	0.007072339
0% Min	0.000898439

Figure 48. QDA resubstitution misses quartiles

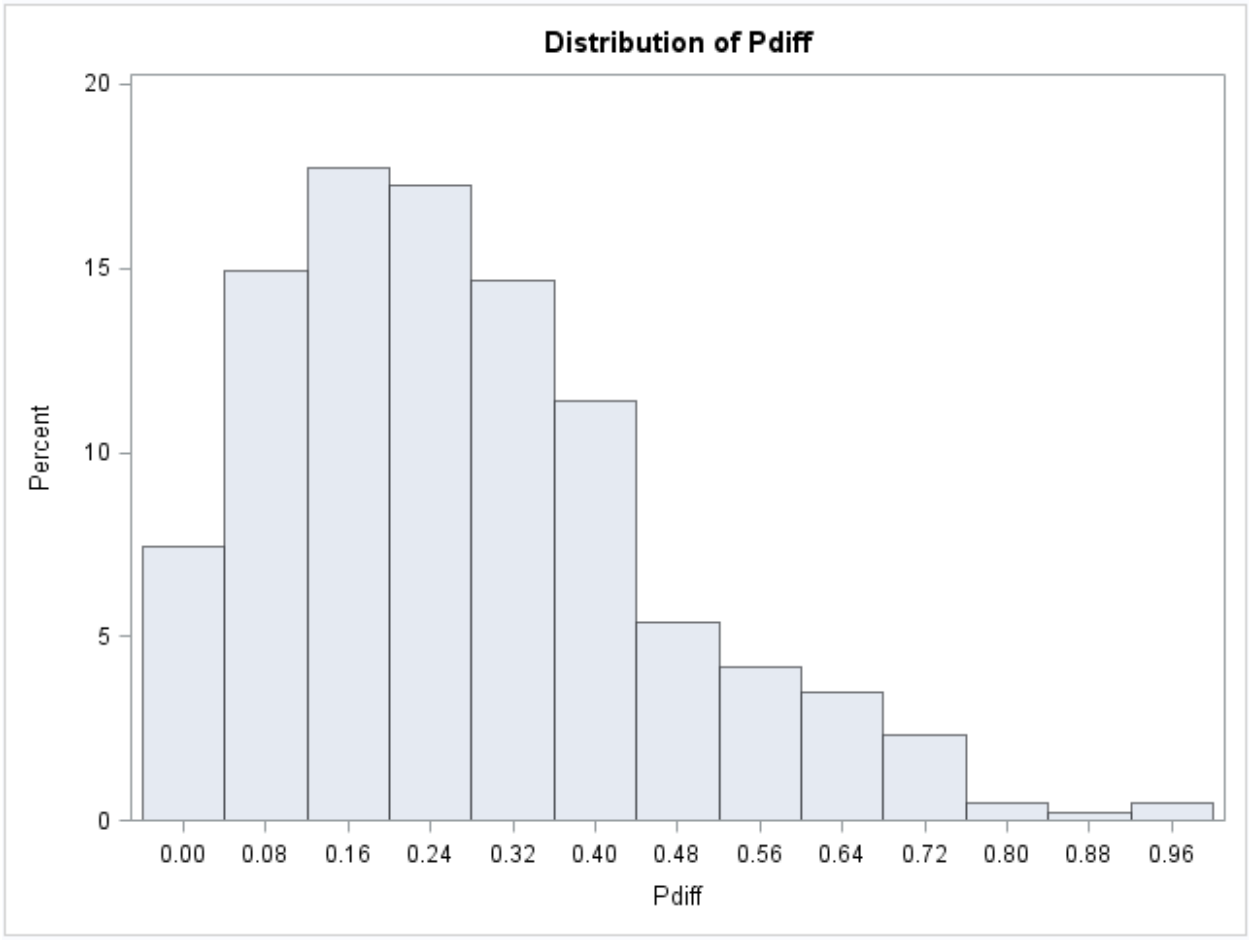

Figure 49. QDA resubstitution misses histogram

Basic Statistical Measures			
Location		Variability	
Mean	0.278586	Std Deviation	0.19131
Median	0.250273	Variance	0.03660
Mode	0.001186	Range	0.99619
		Interquartile Range	0.25736

Figure 50. Descriptive statistics - QDA cross-validation misses

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	0.997035083
99%	0.800448359
95%	0.654495062
90%	0.562619010
75% Q3	0.388082384
50% Median	0.250272885
25% Q1	0.130718679
10%	0.056454702
5%	0.026425974
1%	0.002304943
0% Min	0.000844887

Figure 51. QDA cross-validation quartiles

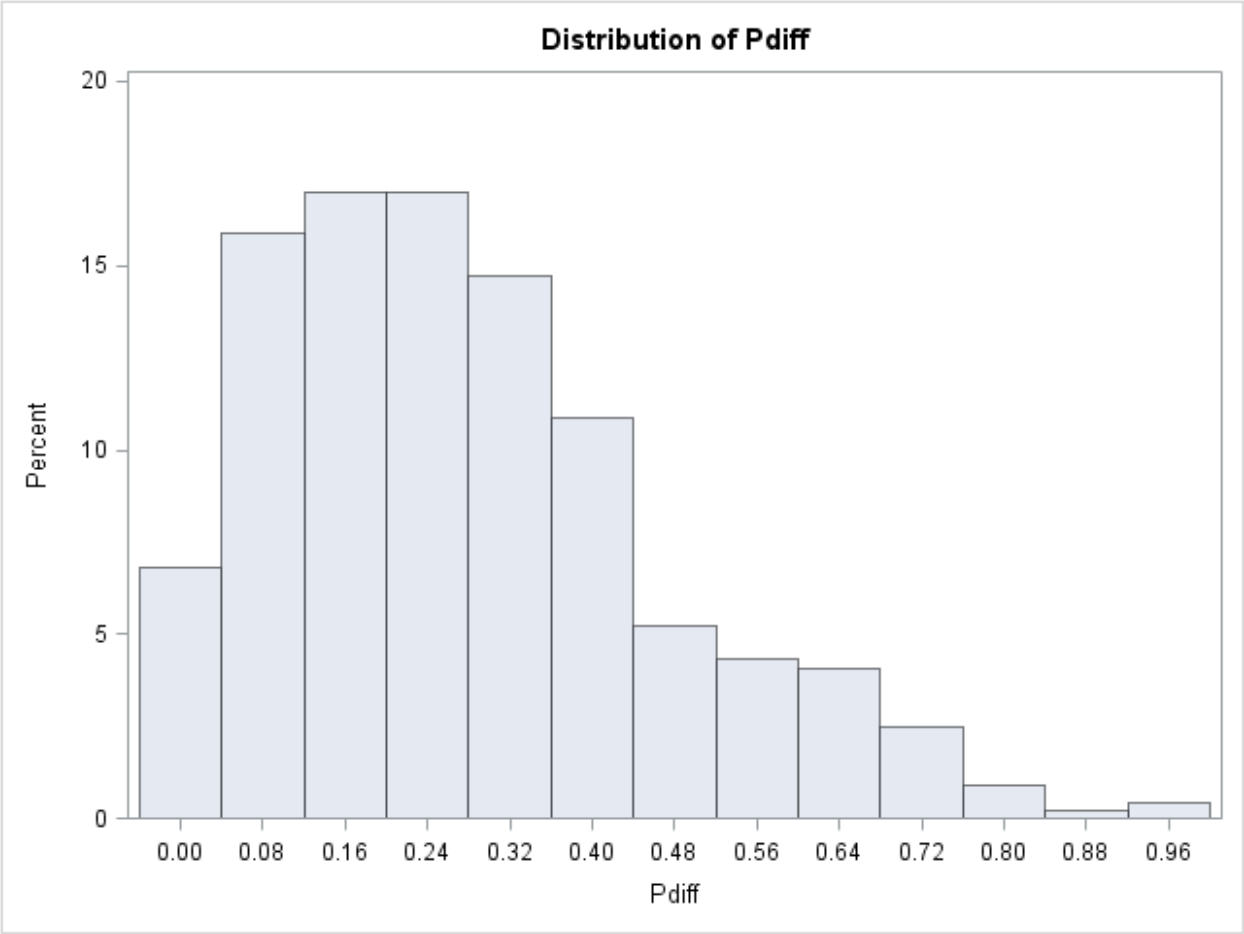

Figure 52. QDA cross-validation histogram

Basic Statistical Measures			
Location		Variability	
Mean	0.289666	Std Deviation	0.19088
Median	0.277028	Variance	0.03644
Mode	0.004812	Range	0.92121
		Interquartile Range	0.25835

Figure 53. Descriptive statistics - QDA holdout misses

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	0.92474635
99%	0.90402326
95%	0.64453678
90%	0.55833531
75% Q3	0.38991430
50% Median	0.27702840
25% Q1	0.13156029
10%	0.05900402
5%	0.03660866
1%	0.00386892
0% Min	0.00353980

Figure 54. QDA holdout misses quartiles

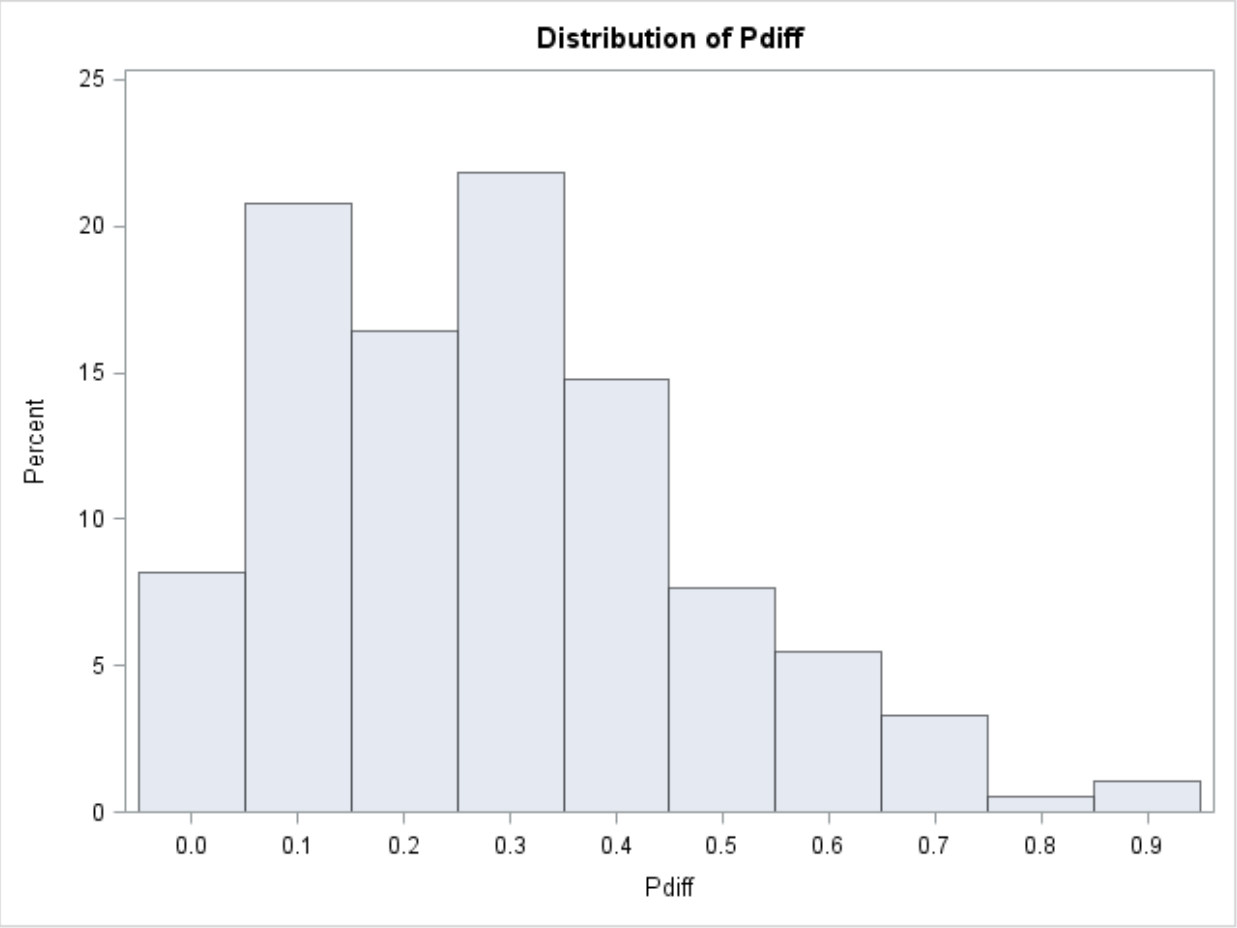

Figure 55. QDA holdout misses histogram

VITA

David Mercer was born in Cincinatti, OH. He began his higher education at Lander University in Greenwood, SC, studying chemistry and chemical engineering. Upon transferring to Clemson University, David changed his major to sociology. In 2010 he graduated from Clemson with a Bachelor of Science in sociology emphasizing in criminal justice and a minor in philosophy. David became interested in forensic anthropology while studying at Clemson. He was admitted to The University of Tennessee Anthropology program. He was offered a graduate research assistantship at the Forensic Anthropology Center, where he worked for two years. In his final year of graduate study he worked for the Office of Information Technology's Research Computing Support group, providing students and professors statistical consulting. He graduated with a Master of Arts in anthropology and a graduate minor in statistics in May 2013.